

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Metodología aplicada al análisis masivo de datos (o Análisis estadístico de datos genómicos).

Ramón Díaz-Uriarte

<http://ligarto.org/rdiaz>

14-02-2008

Objetivos de esta clase

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

- Ser conscientes de que de los datos a las conclusiones biológicas/biomédicas hay un conjunto de pasos que requieren (impepinablemente) estadística.
- Quereis hacer inferencias en un mundo ruidoso.
- Conocer los “grandes temas” en las preguntas que se plantean
- Entender el origen de algunos problemas en el uso de la estadística
- Entender cuándo hay que hablar con un estadístico (siempre —o casi siempre)
- Ser conscientes del tipo de cosas que el estadístico está pensando

Lo que esta clase NO es

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

- Una introducción a la estadística (no hay tiempo)
- Toda la estadística que necesitais para analizar vuestros datos (para eso mucho menos)
- Libro de recetas estadísticas
- El manual de un programa estadístico
- Utilizaremos ejemplos simples. Muuuuuuuuucho más simple que cualquier cosa que jamás analizareis.

Outline

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Microarrays de expresión: preguntas habituales

Hay grupos? Clustering

Expresión diferencial

Clasificación

Qué preguntas se suele intentar querer contestar?

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

- Hay grupos en los genes?
- Hay grupos en los sujetos?
- Hay diferencias en la expresión de ciertos genes entre los grupos de sujetos?
- Existen genes que nos permitan diferenciar entre grupos de pacientes?

Microarrays de
expresión:
preguntas
habituales

Hay grupos? Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Hay grupos? Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión diferencial

Expresión diferencial vs. clasificación

Expresión diferencial: test estadístico

Control de multiple testing

Clasificación

Introducción

Etapas

Estimar error del clasificador

Ultimas observaciones

Hay grupos?

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación

Expresión diferencial: test
estadístico

Control de multiple testing

Clasificación

Introducción

Etapas

Estimar error del
clasificador

Ultimas observaciones

- ¿Podemos encontrar distintos grupos de genes que se comportan de forma parecida y cuyo comportamiento es distinto al de otros grupos de genes?
- ¿Podemos encontrar distintos grupos de sujetos que se comportan de forma parecida y cuyo comportamiento es distinto al de otros grupos de sujetos?

“Class discovery”, clustering, analisis de aglomerados

Sólo tiene sentido si ...

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación

Expresión diferencial: test
estadístico

Control de multiple testing

Clasificación

Introducción

Etapas

Estimar error del
clasificador

Ultimas observaciones

Preguntas que sólo tienen sentido **si no conocemos, de antemano, de la existencia de grupos de sujetos/genes.**

Dos piezas necesarias

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación

Expresión diferencial: test
estadístico

Control de multiple testing

Clasificación

Introducción

Etapas

Estimar error del
clasificador

Ultimas observaciones

Definir qué es “comportarse de forma parecida” y poder medir “distancia”.

Describir y/o definir como agrupamos en función de esas distancias.

Primera pieza: Distancia

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación

Expresión diferencial: test
estadístico

Control de multiple testing

Clasificación

Introducción

Etapas

Estimar error del
clasificador

Ultimas observaciones

- Distancias (por ejemplo, distancia euclídea).
- Correlaciones

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión diferencial

Expresión diferencial vs. clasificación

Expresión diferencial: test estadístico

Control de multiple testing

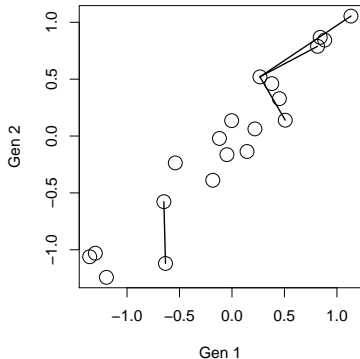
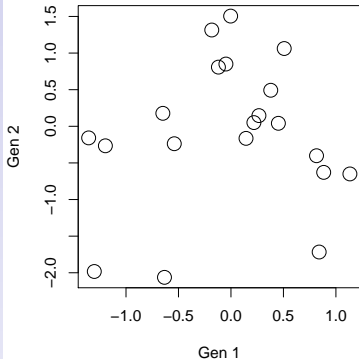
Clasificación

Introducción

Etapas

Estimar error del clasificador

Ultimas observaciones



Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Al final, tenemos una matriz de distancias entre todos los genes, y una matriz de distancias entre todos los sujetos.

¿Y ahora?

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación

Expresión diferencial: test
estadístico

Control de multiple testing

Clasificación

Introducción

Etapas

Estimar error del
clasificador

Ultimas observaciones

	s1	s2	s3	s4
s1	-	2	7	3
s2	-	-	8	4
s3	-	-	-	9
s4	-	-	-	-

???

Segunda pieza: Algoritmos de agrupación

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación

Expresión diferencial: test
estadístico

Control de multiple testing

Clasificación

Introducción

Etapas

Estimar error del
clasificador

Ultimas observaciones

- Jerárquicos:
 - ▶ Divisivos
 - ▶ Aglomerativos
- No jerárquicos (especificar número de clusters).

Jeraquicos (e.g., aglomerativos)

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- Juntar los dos que tengan menor distancia (i.e., estatura mas parecida).
- Continuar juntando, hasta que todas las muestras (todos los sujetos) en algún grupo.

Jeraquicos (e.g., aglomerativos)

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- Juntar los dos que tengan menor distancia (i.e., estatura mas parecida).
- Continuar juntando, hasta que todas las muestras (todos los sujetos) en algún grupo.
- ¿Cómo continuar juntando? La nueva muestra, ¿a quien se tiene que parecer?

No jerárquicos

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- Sospechamos que existen dos grupos.
- Encontrar la asignación de todos los elementos a dos grupos de forma que “sea la mejor solución”. Por ejemplo: la suma de distancias de cada observación a su “centro del cluster” sea mínima..
- (La matriz de distancias entre puntos no nos hace falta; sí, en este caso, de los puntos al centro del cluster).

Problemitas ...

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación

Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- ¿Cuál es la medida de distancia apropiada?
- ¿Cuál es el algoritmo apropiado?
- ¿Queremos usar todos los genes cuando agrupamos sujetos?

Precauciones

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- El clustering es “class discovery”: es una herramienta exploratoria, NO una herramienta confirmatoria (con alguna excepción).
- El clustering SIEMPRE devuelve clusters, haya o no estructura en los datos.
- Que un cluster sea “relevante”, “estable” es una pregunta distinta.
- Clustering no es la herramienta apropiada si conocemos de antemano la asignación a grupos.

Microarrays de expresión: preguntas habituales

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

Hay grupos? Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión diferencial

Expresión diferencial vs. clasificación

Expresión diferencial: test estadístico

Control de multiple testing

Clasificación

Introducción

Etapas

Estimar error del clasificador

Ultimas observaciones

¿Hay diferencias en la expresión de ciertos genes entre los grupos de sujetos?

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

Si tenemos 2 (o 3, o 4, o ...) tipos de sujetos (cáncer de mama, cáncer de colón, etc), ¿qué genes muestran expresión diferencial?

¿Hay diferencias en la expresión de ciertos genes entre los grupos de sujetos?

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

Si tenemos 2 (o 3, o 4, o ...) tipos de sujetos (cáncer de mama, cáncer de colón, etc), ¿qué genes muestran expresión diferencial?

Dados dos (o tres, o cuatro, o ...) tipos de sujetos, ¿qué genes hacen cosas distintas?

Y esto, ¿en qué se diferencia de nuestra cuarta pregunta?

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación

Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

¿Existen genes que nos permitan diferenciar entre grupos de pacientes?

(vs. ¿qué genes muestran diferencias entre grupos de sujetos?)

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

**Expresión diferencial vs.
clasificación**
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

La estatura es distinta entre hombres y mujeres españoles.

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

La estatura es distinta entre hombres y mujeres españoles.

La estatura es muy mala para distinguir: sujeto X mide 1.74, ¿es hombre o mujer?

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

**Expresión diferencial vs.
clasificación**

Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

La relación entre cantidad de grasa en caderas y hombros...

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

La relación entre cantidad de grasa en caderas y
hombros...

podría servir para distinguir, aunque la cantidad en cada
uno, individualmente, no sirva para mucho a la hora de
distinguir.

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

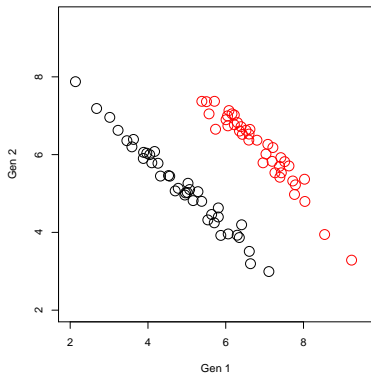
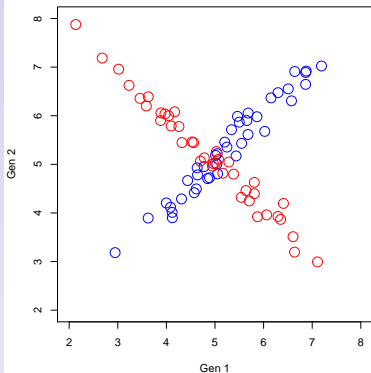
Expresión diferencial

Expresión diferencial vs. clasificación

Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones



Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
**Expresión diferencial: test
estadístico**
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

... todos habeis hecho bioestadística en alguna vida
pasada ...

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
**Expresión diferencial: test
estadístico**
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

... todos habeis hecho bioestadística en alguna vida
pasada ...

¿Hace falta repasar que es un estadístico y un p-valor?

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
**Expresión diferencial: test
estadístico**
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Queremos comparar la media de expresión del gen MYC entre 10 pacientes con cáncer de mama y 12 pacientes sanas. ¿Cómo?

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
**Expresión diferencial: test
estadístico**
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Queremos comparar la media de expresión del gen MYC entre 10 pacientes con cáncer de mama y 12 pacientes sanas. ¿Cómo?

Más formalmente: ¿puede la “verdadera” (media de la) expresión en los dos grupos ser igual? (¿Tienen los dos grupos la misma media de expresión?)

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
**Expresión diferencial: test
estadístico**
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Queremos comparar la media de expresión del gen MYC entre 10 pacientes con cáncer de mama y 12 pacientes sanas. ¿Cómo?

Más formalmente: ¿puede la “verdadera” (media de la) expresión en los dos grupos ser igual? (¿Tienen los dos grupos la misma media de expresión?)

Mejor aun si decimos algo sobre la certeza en la conclusión de “son iguales” o “son distintas”.

Calculamos la media en los dos grupos: 2.2 y 3.4. ¿Y?

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
**Expresión diferencial: test
estadístico**
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Calculamos la media en los dos grupos: 2.2 y 3.4. ¿Y?

La diferencia es 1.2. ¿Es esa diferencia mucha o poca?

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
**Expresión diferencial: test
estadístico**
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Calculamos la media en los dos grupos: 2.2 y 3.4. ¿Y?

La diferencia es 1.2. ¿Es esa diferencia mucha o poca?

La media de expresión del gen XYZ, cuando calculamos la media en muestras como las de arriba, tiene el 90% de sus valores entre 1.1 y 1.12. Una diferencia de 1.2 es muuuuucho.

La media de expresión del gen UTV, . . . , tiene el 90% de sus valores entre 1.1 y 8.2. Una diferencia de 1.2 es pooooooco.

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

Calculamos la media en los dos grupos: 2.2 y 3.4. ¿Y?

La diferencia es 1.2. ¿Es esa diferencia mucha o poca?

La media de expresión del gen XYZ, cuando calculamos la media en muestras como las de arriba, tiene el 90% de sus valores entre 1.1 y 1.12. Una diferencia de 1.2 es muuuuucho.

La media de expresión del gen UTV, . . . , tiene el 90% de sus valores entre 1.1 y 8.2. Una diferencia de 1.2 es pooooooco.

“Como de relevante” es una diferencia depende de la variabilidad en la diferencia de las medias.

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

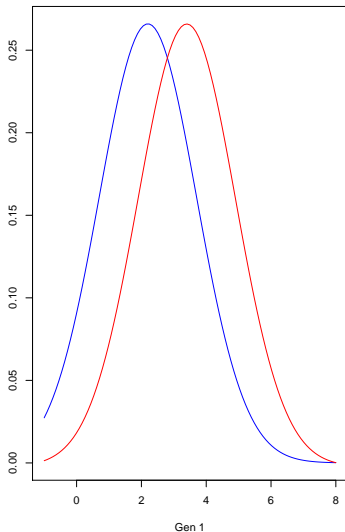
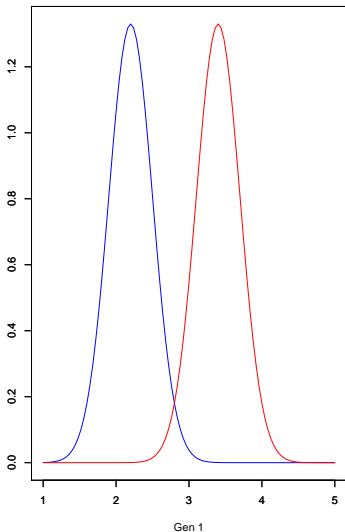
Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones



Para comparar dos grupos

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

1. Calcular las medias
2. Restarlas
3. Calcular una cantidad relacionada con la varianza de la diferencia de medias (esa cantidad se calcula a partir de la varianza de cada grupo).
4. Dividir la diferencia de medias por la desviación típica de la diferencia de las medias.
5. Ya tenemos una “diferencia estandarizada”: el estadístico de la t.

¿Y esa medida de probabilidad?

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

1. Usando distintas posibles estrategias (análisis, permutación) podemos obtener la distribución de “t” bajo la hipótesis nula.
2. Hipótesis nula en este caso: las dos medias de verdad son iguales.
3. Obtener la distribución de los “t” que uno calcula si, en realidad, no hay diferencias.
4. Calculamos la probabilidad de observar nuestro “t” si la hipótesis nula es cierta.
5. p-valor: cómo de probable nuestro resultado si la nula fuera cierta.
6. p-valor: medida de evidencia contra la hipótesis nula.

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

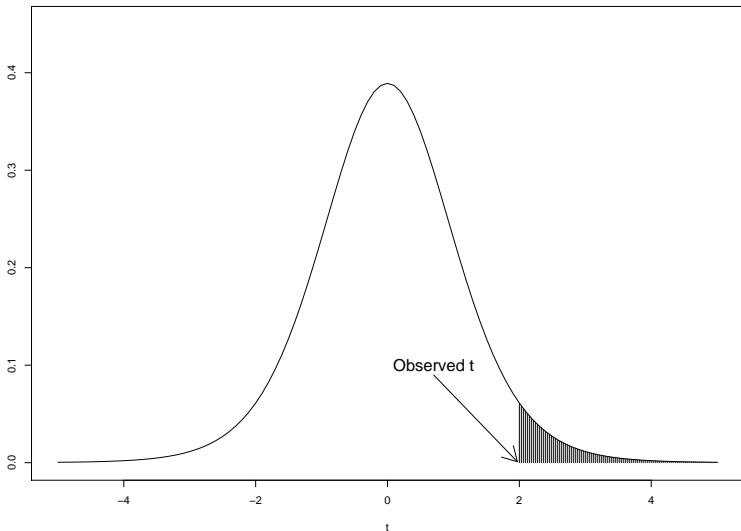
Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones



De vuelta a las arrays

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Ya sabemos como obtener un p-valor para comparar dos grupos.

(Y existen mecanismos similares para otras comparaciones, entre más grupos, o relación con la supervivencia, etc).

¿Podemos simplemente calcular un p-valor por gen y seleccionar aquellos relevantes?

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación

Expresión diferencial: test
estadístico

Control de multiple testing

Clasificación

Introducción

Etapas

Estimar error del
clasificador

Ultimas observaciones

NO

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

NO

No estamos obteniendo el p-valor de un test (un contraste de hipótesis) sino el de miles de tests.

Los peces

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- Nos vamos de pesca.
- En este mar hay un pez concreto (pez A) con una probabilidad de ser pescado de 0.05.
- En ese mar 1000 peces como el A (pero sólo un es A, claro).
- ¿Cuál es $Pr\{\text{cenamos pez A}\}$?
- ¿Cuál es $Pr\{\text{cenamos pescado}\}$?

Los peces (II)

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- $Pr\{\text{cenamos pez } A\} = 0.05.$
- $Pr\{\text{cenamos pescado}\} \simeq 1 .$
- Los eventos “cenarnos al pez A” y “cenar pescado” son muy diferentes.
- $\text{Cenar pescado} = \bigcup(\text{cenarnos a } A, \text{cenarnos a } B, \text{cenarnos a } C, \dots, \text{cenarnos a } A \text{ y } B, \dots).$

Los p-values son peces

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- Si tenemos 30000 genes, en los que no existen ninguna diferencia . . .
- y declaramos como “interesantes” todos los que tienen $p - value < 0.05$ vamos a cometer montones de “falsos positivos” (~ 1500).
- Necesitamos controlar eso.

The p-value case

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

(An example modified from Westfall and Young, 1993 “Resampling-based multiple testing”).

- Suppose we have 10 independent genes. Thus, 10 null hypotheses, one for each gene.
- Suppose also that there are no differences in gene expression between the two groups of patients (i.e., the null is true, and we are using the appropriate test so that the p-value is Uniform on $[0,1]$).
- Thus, the probability that a particular test (say, for gene 3) is declared significant at level 0.05 is exactly 0.05. Good.

p-value case (II)

- However, the probability of declaring at least one of the 10 hypotheses false (i.e., rejecting at least one, or finding at least one result significant) is:

$$\begin{aligned} Pr(\text{at least one null rejected}) &= 1 - Pr(\text{all } p_i > 0.05) = \\ &= 1 - Pr(1 - 0.05)^{10} = 1 - 0.95^{10} = 0.401 \end{aligned}$$

- So now, even if the 10 genes are not differentially expressed, there is a probability of 0.401 (yes, that is 40%!!!) of “finding” at least one which we declare as significantly different.
- The more genes, the more serious is the problem.
- In summary, without control for multiple testing, we would end up rejecting the null much more often than we should.

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

	# no rechazadas	# rechazadas
# verdaderas nulas	U	V
# no-nulas (difs.)	T	S

FDR False Discovery Rate: tasa de descubrimiento falso: proporción esperada de errores de tipo I entre las nulas rechazadas: $(V + S)$. $FDR = E(Q)$ donde $Q = V / (V + S)$ si $V + S > 0$ (y $Q = 0$ en el otro caso).

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

- Tamaño de muestra
- Test apropiado para el problema.
- Test y análisis apropiado al tipo de diseño.

Tamaño de muestra

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- Selecciono al azar 2 varones y 3 mujeres de esta clase. Dinero medio en el bolsillo: 3 euros los varones, 15 euros las mujeres.
- No hace falta un p-valor: el tamaño de muestra es ridículamente pequeño para lo que queremos.

Tamaño de muestra

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- Resultados significativos (o incluso “reales”) vs. resultados repetibles.
- Cada estudio mal hecho es una oportunidad mal aprovechada.
- El argumento del dinero y la analogía del SSC.
- 50 muestras por grupo.

Test apropiado

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

- Incluso para comparar dos muestras independientes hay una variedad de tests.
- ¿Y si hay más grupos?
- ¿Y si hay información sobre variables clínicas?
- ¿Y si los sujetos parcialmente relacionados —parentesco, comunidad autónoma, etc?
- ¿Y si datos de supervivencia?

Microarrays de expresión: preguntas habituales

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

Hay grupos? Clustering

Hay grupos?

Dos piezas necesarias

Medidas de distancia

Algoritmos de agrupación

Problemas

Expresión diferencial

Expresión diferencial vs. clasificación

Expresión diferencial: test estadístico

Control de multiple testing

Clasificación

Introducción

Etapas

Estimar error del clasificador

Ultimas observaciones

Diferenciar entre grupos de pacientes

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción

Etapas
Estimar error del clasificador
Ultimas observaciones

Clasificación (o predicción si variable continua o supervivencia).

Un problema clásico en estadística y machine learning. Bastante bien entendido. Y con soluciones estándar y “out of the box”.

¿Qué queremos? Un buen clasificador que, dado una nueva muestra, la ponga en la caja apropiada.

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción

Etapas
Estimar error del
clasificador
Ultimas observaciones

Tenemos muchos más genes que sujetos: muchas más variables que muestras ($p \gg n$).

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción

Etapas
Estimar error del
clasificador
Ultimas observaciones

Tenemos muchos más genes que sujetos: muchas más variables que muestras ($p \gg n$).

Esto es “el mundo al revés”.

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador
Ultimas observaciones

Tenemos muchos más genes que sujetos: muchas más variables que muestras ($p \gg n$).

Esto es “el mundo al revés”.

Y nos sobra información redundante.

Ideas clave

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción

Etapas
Estimar error del
clasificador
Ultimas observaciones

- Todo lo que nos importa es obtener un buen clasificador.
- Los p-valores nos dan igual.
- Tendremos que seleccionar algunos genes.
- Tendremos, **MUY ESPECIALMENTE**, que estimar el error del clasificador.

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Ultimas observaciones

- Selección de un algoritmo de clasificación.
- Selección de genes.
- Construcción del clasificador.
- Estimar error del clasificador.

Estimar el error del clasificador

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
**Estimar error del
clasificador**
Ultimas observaciones

Muestra de 50 sujetos con cáncer y 50 sin cáncer.
Construimos nuestro algoritmo con esas 100 muestras, y
en esa muestra de 100 cometemos un error del 10%.

Estimar el error del clasificador

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
**Estimar error del
clasificador**

Ultimas observaciones

Muestra de 50 sujetos con cáncer y 50 sin cáncer.
Construimos nuestro algoritmo con esas 100 muestras, y
en esa muestra de 100 cometemos un error del 10%.

¿Podemos usar ese 10% como una estimación razonable
del error que cometeríamos con unas nuevas muestras?

Validación cruzada

- Supongamos 100 sujetos, 50 cáncer y 50 no cáncer.
- Seleccionar al azar 10 (“testing set”).
- Usar los otros 90 para construir el clasificador (“training set”).
- Evaluar el clasificador en los 10 primeros.
- Repetir este proceso otras 9 veces (hasta que todos los sujetos hayan sido usados exactamente una vez en el “testing set”).
- Tenemos 10 estimaciones de error, calculamos la media, y tenemos ahora una estimación (más o menos) insesgada del error que cometeríamos con una nueva muestra.

Microarrays de expresión:
preguntas habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión diferencial

Expresión diferencial vs. clasificación
Expresión diferencial: test estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del clasificador
Últimas observaciones

Ojo con el “selection bias”

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
**Estimar error del
clasificador**
Ultimas observaciones

¿Y si hemos hecho selección de genes?

- Seleccionamos los 100 genes con mejor p-valor.
- Construimos clasificador

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
**Estimar error del
clasificador**
Ultimas observaciones

El proceso de validación cruzada ha de incorporar la selección de genes.

Hay que hacer la selección en cada uno de los subgrupos de “entrenamiento”.

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
**Estimar error del
clasificador**
Ultimas observaciones

- Existen otras técnicas relacionadas con la validación cruzada, como el bootstrap, etc.
- En cualquier caso, el dejar aparte un sólo conjunto de testeo es una muy mala idea.

Microarrays de
expresión:
preguntas
habituales

Hay grupos?
Clustering

Hay grupos?
Dos piezas necesarias
Medidas de distancia
Algoritmos de agrupación
Problemas

Expresión
diferencial

Expresión diferencial vs.
clasificación
Expresión diferencial: test
estadístico
Control de multiple testing

Clasificación

Introducción
Etapas
Estimar error del
clasificador

Últimas observaciones

- Muchos métodos razonables soluciones similares, incluidos métodos razonables pero bien sencillos (DLDA, KNN).
- Inestabilidad y multiplicidad en soluciones.
- Cual es el mejor número de genes es difícil de determinar.
- ¿Para qué hacemos esto? Interpretación biológica o desarrollo de herramientas diagnósticas.