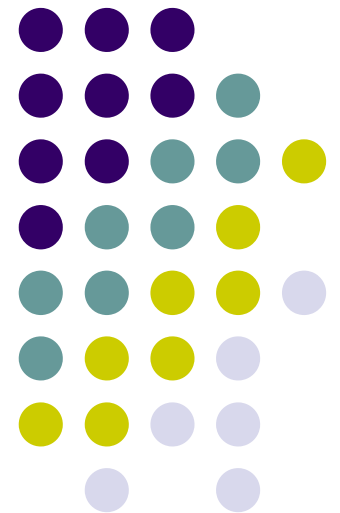


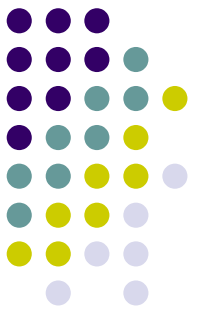
# Selección de variables

---

Basado en parte en curso de I. Guyon

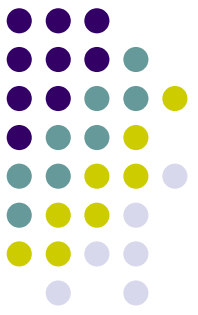


# Selección de variables



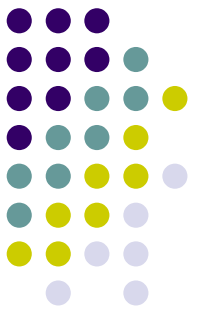
- Por qué y para qué?
- Métodos
- Filtros
- Wrappers
- RFE
- Estabilidad. Selección en listas múltiples

# Selección de variables: Por qué?



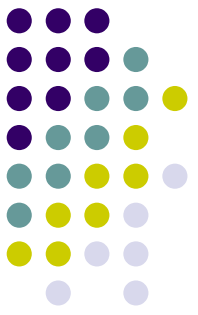
- Muchos problemas actuales tienen cientos o miles de variables medidas (sobre pocos ejemplos)
- Modelar esos problemas “directamente” suele ser sub-óptimo.
  - Tanto en calidad como en interpretabilidad.
- En algunos casos la “extracción de variables” (pca, ica, etc.) no es una opción válida.

# Selección de variables: Para qué?



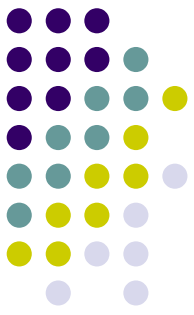
- Para mejorar la performance de los métodos de aprendizaje:
  - Algunos métodos trabajan mucho mejor con menos variables.
    - Aunque los métodos modernos de ML suelen ser muy resistentes al problema de la dimensionalidad.
  - En ciertos casos muchas variables no son informativas del problema (ruido o redundancias).
    - Al eliminarlas reducimos el riesgo de sobreajuste.

# Selección de variables: Para qué?

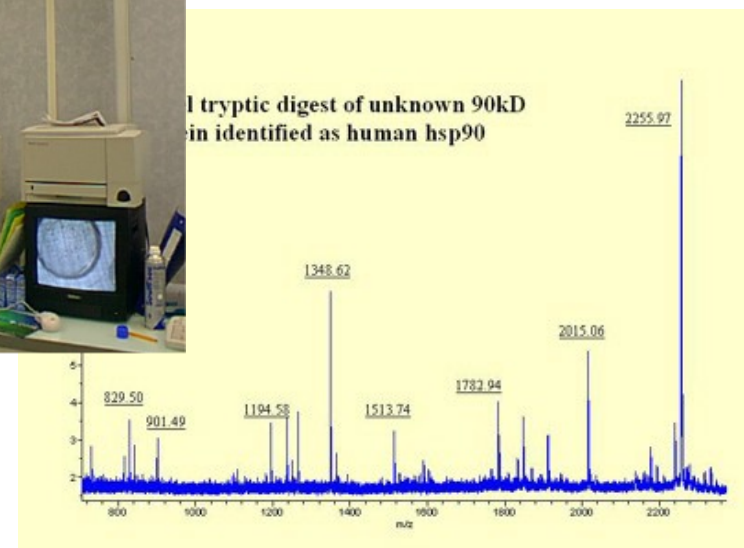
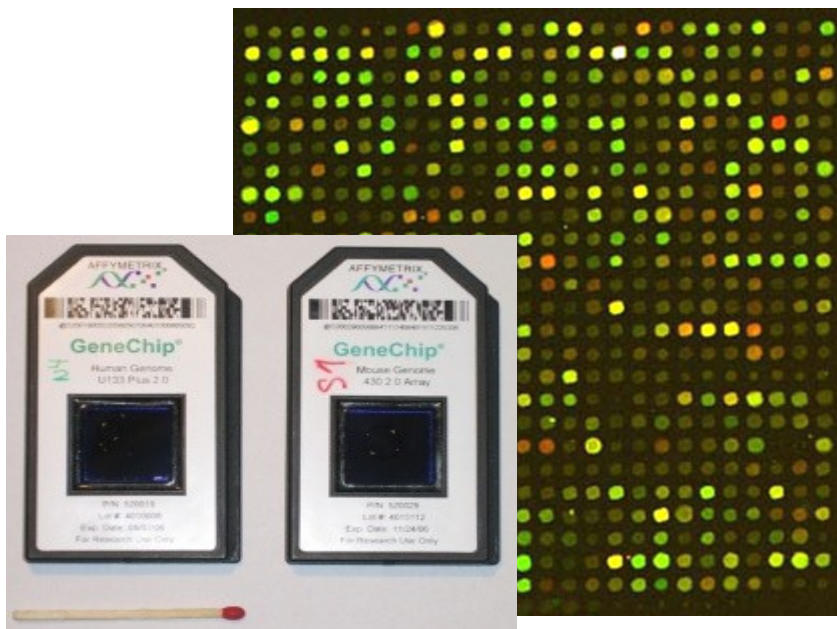


- Para descubrir:
  - Cuáles son las variables más importantes en un problema.
  - Cuáles variables están correlacionadas, co-reguladas, o son dependientes y cuáles no.
- La selección de variables no es más una técnica de pre-procesado, actualmente es una herramienta para descubrir información de un problema.

# Ejemplos actuales



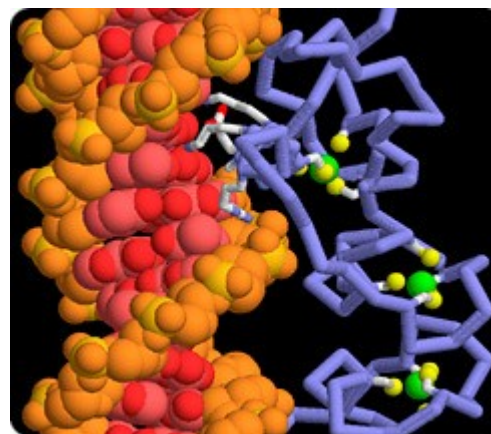
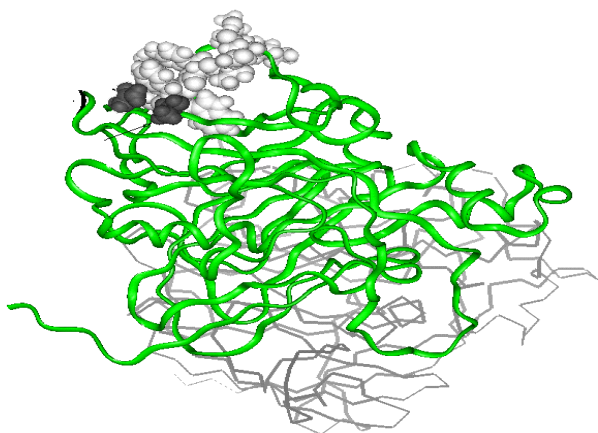
- Técnicas biológicas de “High throughput”
  - DNA Microchips (3000~12000 genes)
  - Mass Spectrometry (200~10000 picos)
  - Nunca más de ~100 muestras.



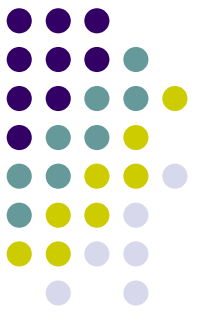
# Ejemplos actuales



- QSAR
  - Relación cuantitativa entre estructura molecular y actividad del compuesto. Clave en la industria farmacéutica.
  - (100~2000) descriptores moleculares.



# Métodos



- **Problema Base:**

Seleccionar un subconjunto óptimo de ( $r$ ) variables de las  $p$  variables originales, dado un criterio.

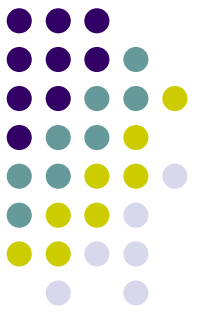
- Por qué no evaluar todas las posibilidades?

Explosión combinatoria: 
$$\sum_{r=1}^p C_r^p = \sum_{r=1}^p \frac{p!}{r!(p-r)!}$$

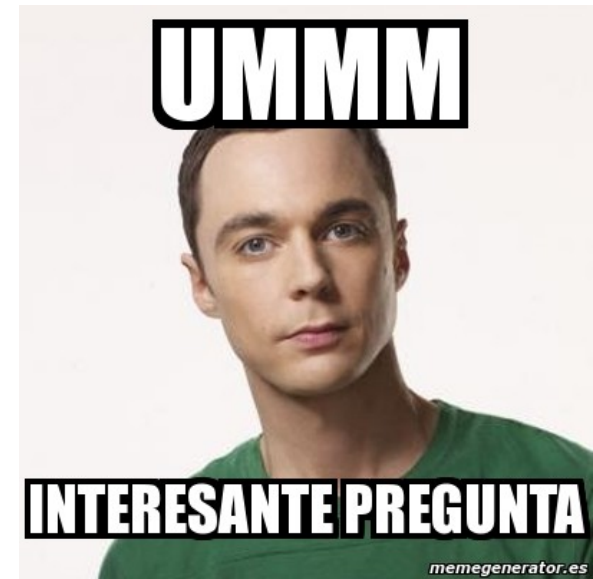
Se usan soluciones sub-óptimas sobre eurísticas.

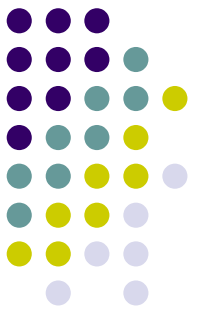


# Cómo elegir variables?



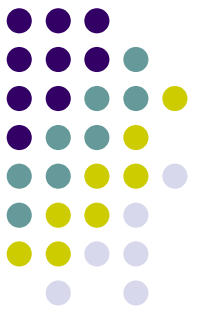
- Qué cosas sería razonable intentar?





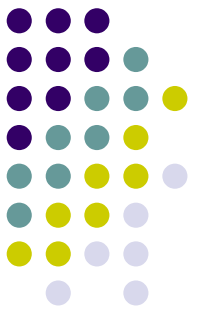
# Métodos

- **Univariados** consideran una variable a la vez.
- **Multivariados:** consideran subconjuntos de variables al mismo tiempo.
- **Filtros:** Ordenan las variables con criterios de importancia independientes del predictor.
- **Wrappers:** Usan el predictor final para evaluar la utilidad de las variables.



# Métodos

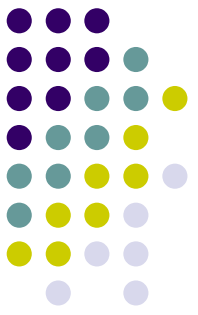
- **Univariados** consideran una variable a la vez.
- **Multivariados:** consideran subconjuntos de variables al mismo tiempo.
- **Filtros:** Ordenan las variables con criterios de importancia independientes del predictor.
- **Wrappers:** Usan el predictor final para evaluar la utilidad de las variables.



# Métodos

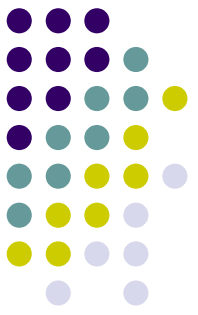
- **Univariados** consideran una variable a la vez.
- **Multivariados:** consideran subconjuntos de variables al mismo tiempo.
- **Filtros:** Ordenan las variables con criterios de importancia independientes del predictor.
- **Wrappers:** Usan el predictor final para evaluar la utilidad de las variables.

# Métodos de Filtro



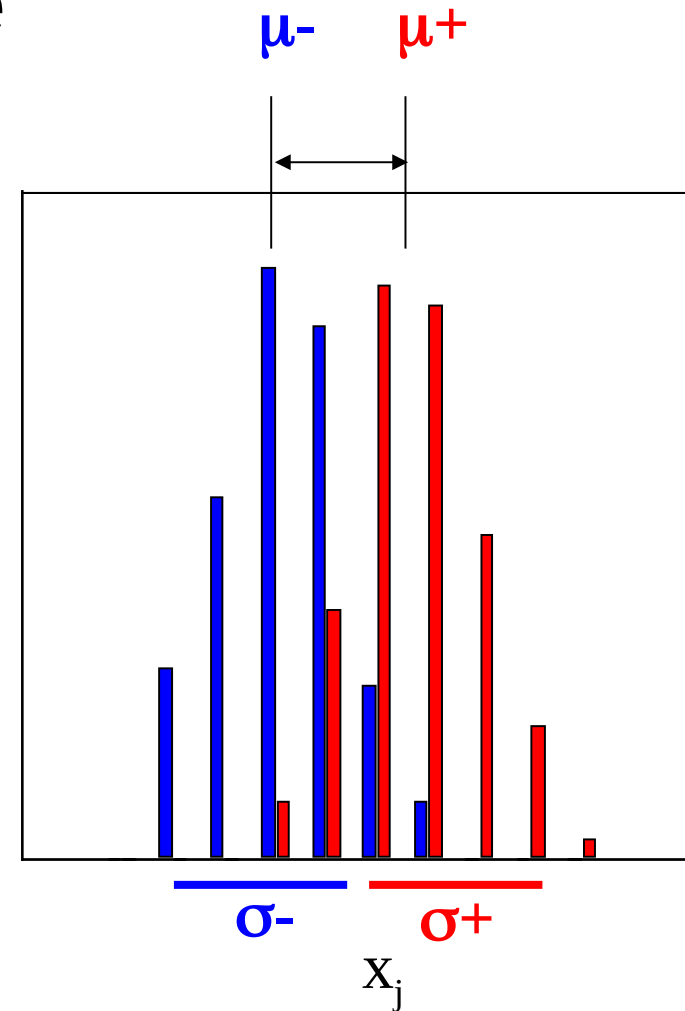
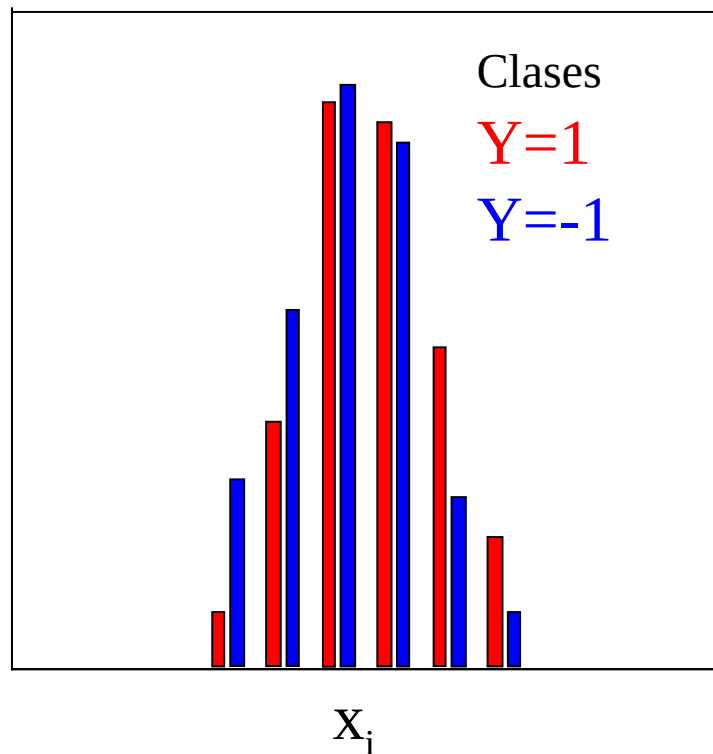
- Elige las mejores variables usando criterios razonables de “importancia”.
- El criterio es preferiblemente independiente del que se usará para modelar.
- Usualmente se usan criterios univariados.
- Se ordenan las variables en base al criterio y se retienen las más importantes (criterio de corte!)

# Métodos de Filtro: ejemplos

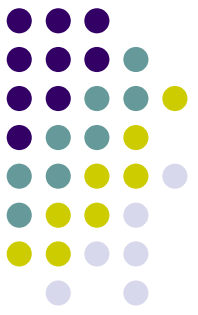


- Clasificación: Relevantes e Irrelevantes

densidad

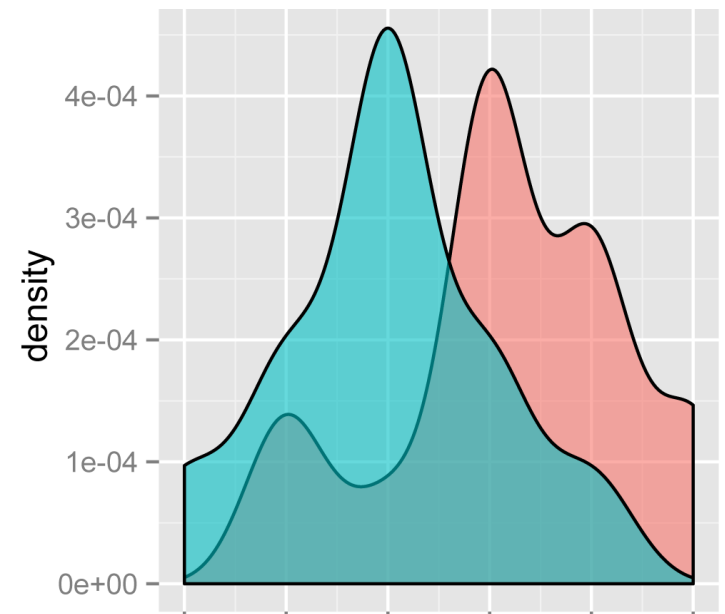
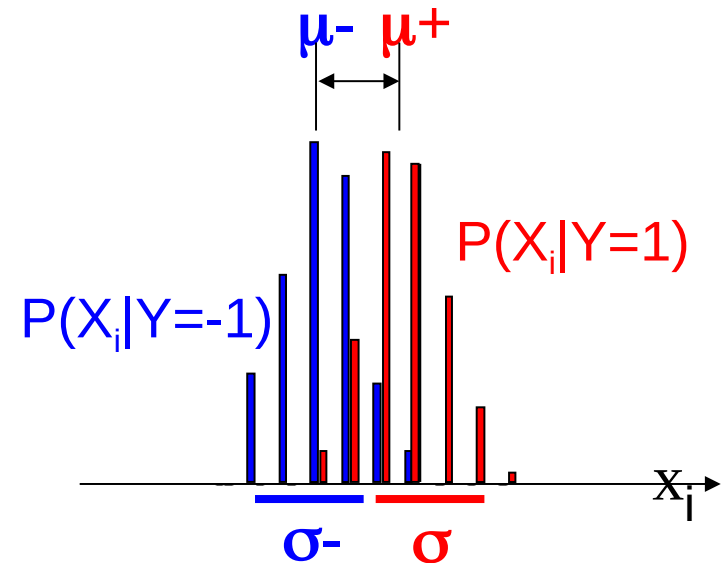


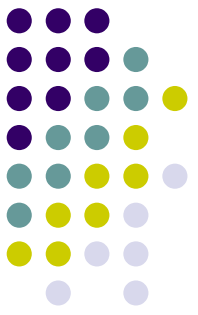
# Métodos de Filtro: ejemplos de criterios



- Clasificación:

- ANOVA: significancia de un t-test entre las clases dada la variable.
- Ganancia de información sobre la clase dada la variable
- Muchos otros (Relief!)
- Kruskal-Wallis: Test no-paramétrico. No asume una distribución. Más general pero menos potente.





# Métodos de Filtro: ejemplos

R:

```
data(iris)
```

```
y<-iris[,5]
```

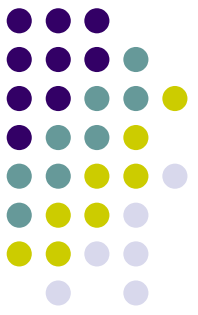
```
#anova – F statistic
```

```
for (i in 1:4){x<-iris[,i];print(oneway.test(x~y)  
  $statistic)}
```

```
#no paramétrico - Kruskal-Wallis
```

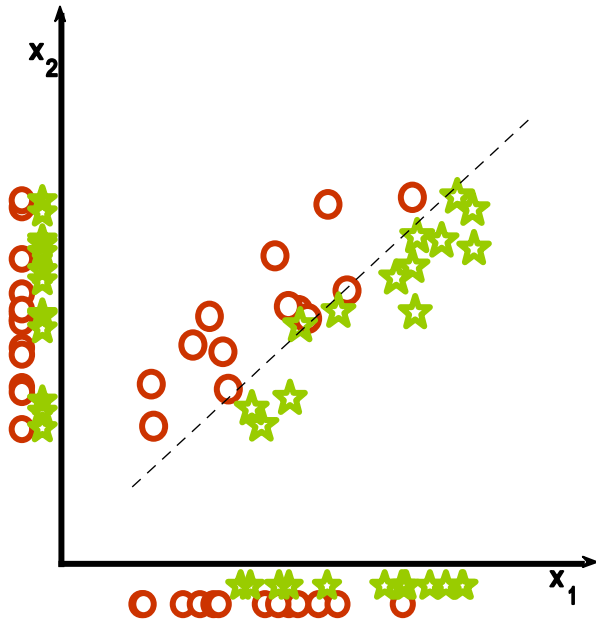
```
for (i in 1:4){x<-iris[,i];print(kruskal.test(x,y)  
  $statistic)}
```





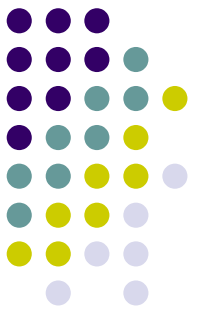
# Multivariados

- Los métodos univariados no pueden resolver algunos problemas

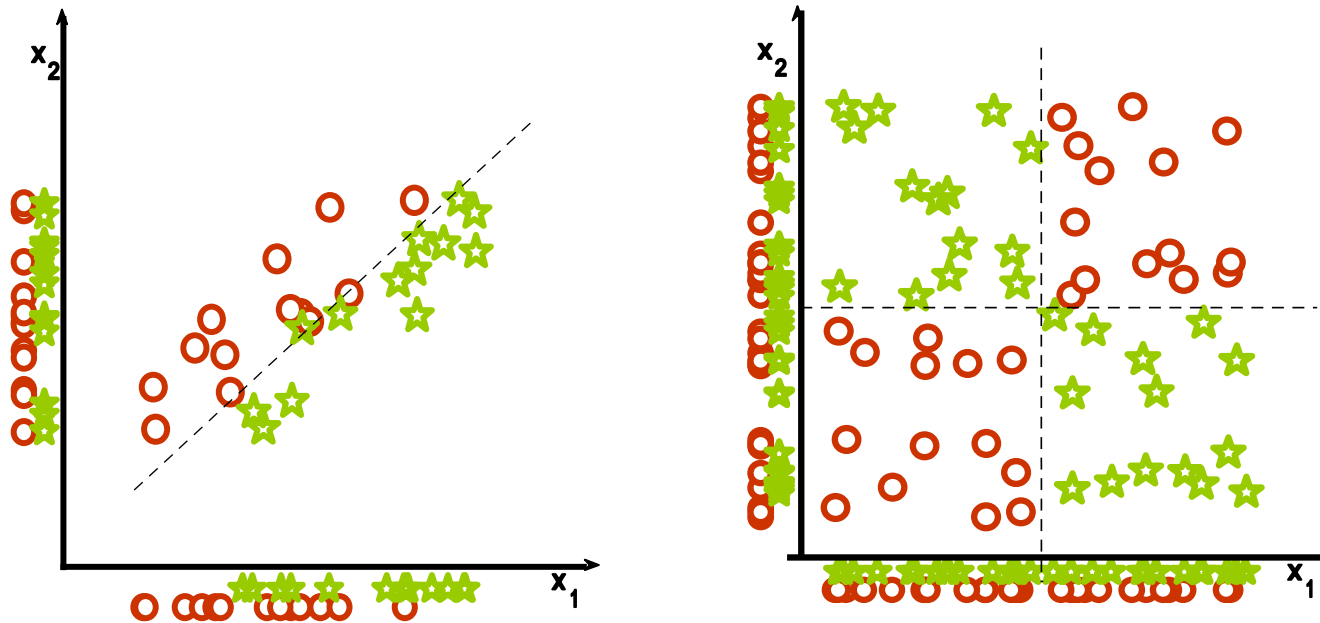


*Guyon-Elisseff, JMLR 2004; Springer 2006*

# Multivariados

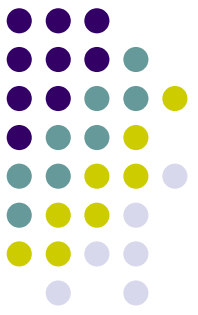


- Los métodos univariados no pueden resolver algunos problemas



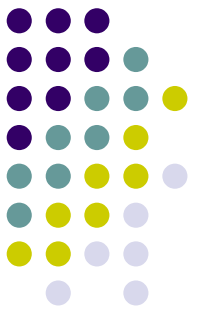
*Guyon-Elisseff, JMLR 2004; Springer 2006*

# Wrappers. Claves



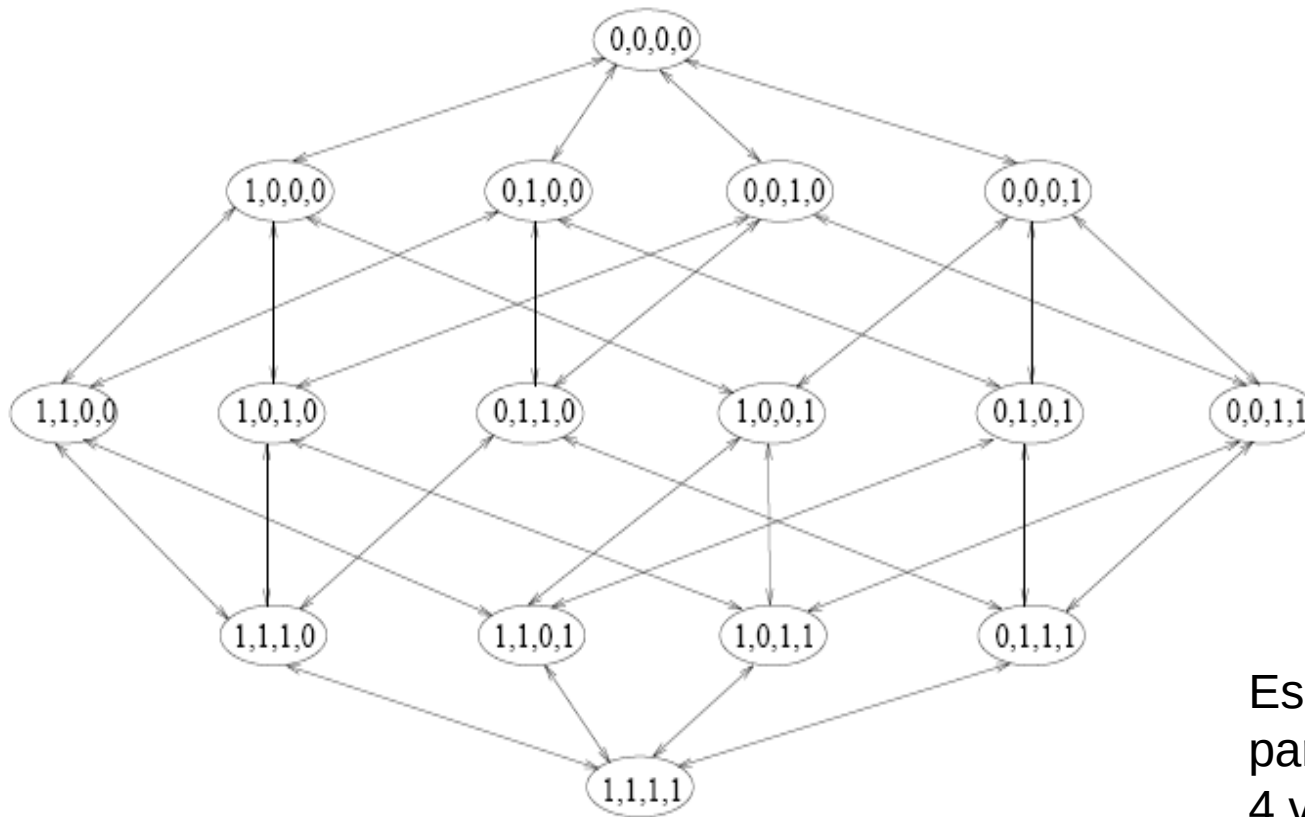
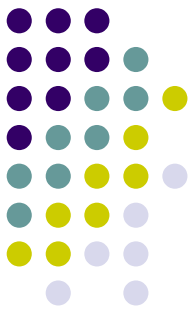
- Seleccionar las mejores variables para modelar (usando el criterio final)
- Para cada subconjunto de variables resolver el problema de modelado. Conservar la mejor solución.
- Como ya discutimos, la búsqueda completa es exponencialmente larga.

# Wrappers. Alternativas



- Búsquedas Greedy:
  - forward selection
  - backward elimination
  - combinaciones de ambas
- Búsquedas pseudo-random:
  - Simulated annealing
  - genetic algorithm

# Wrappers. Ejemplo

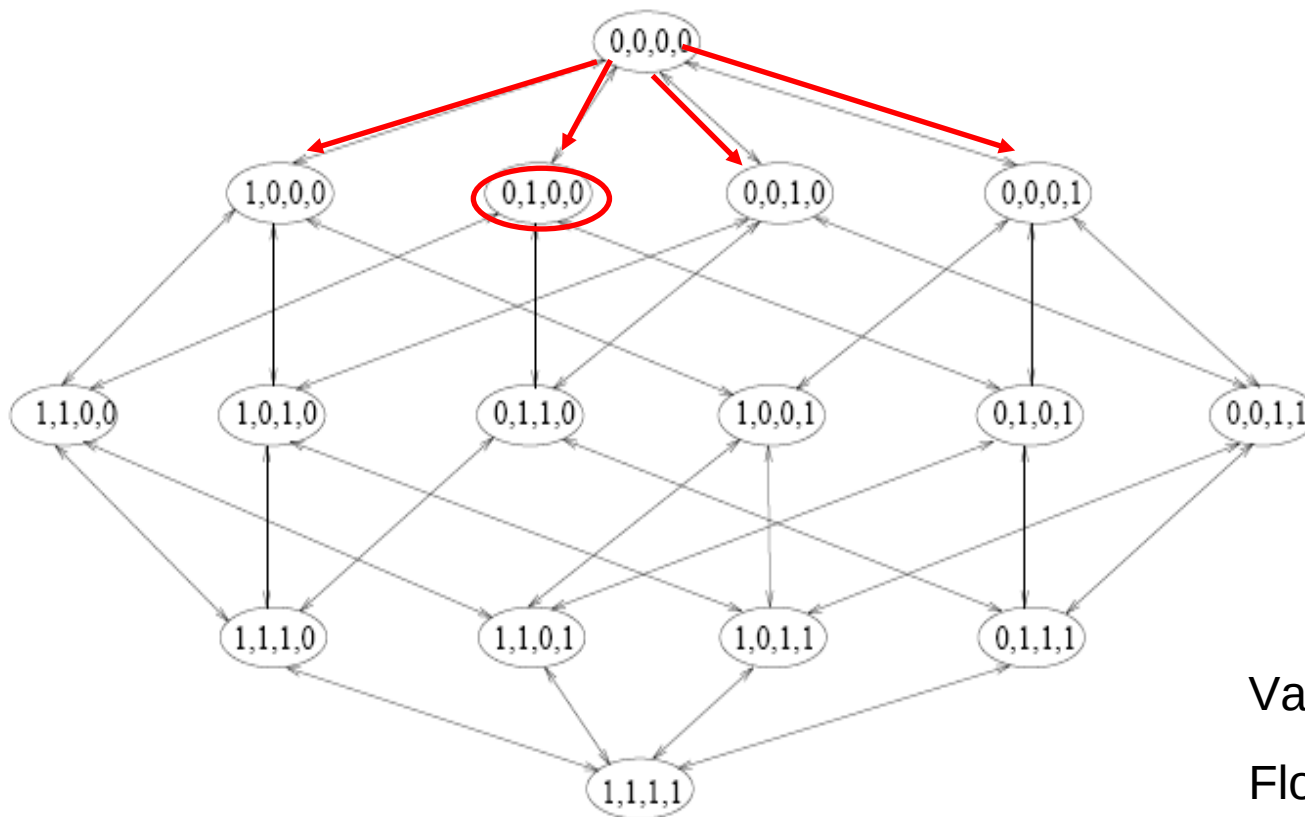
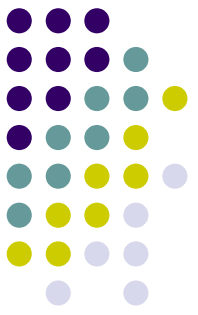


Espacio de búsqueda  
para un problema con  
4 variables.

0 ausente - 1 presente

*Kohavi-John, 1997*

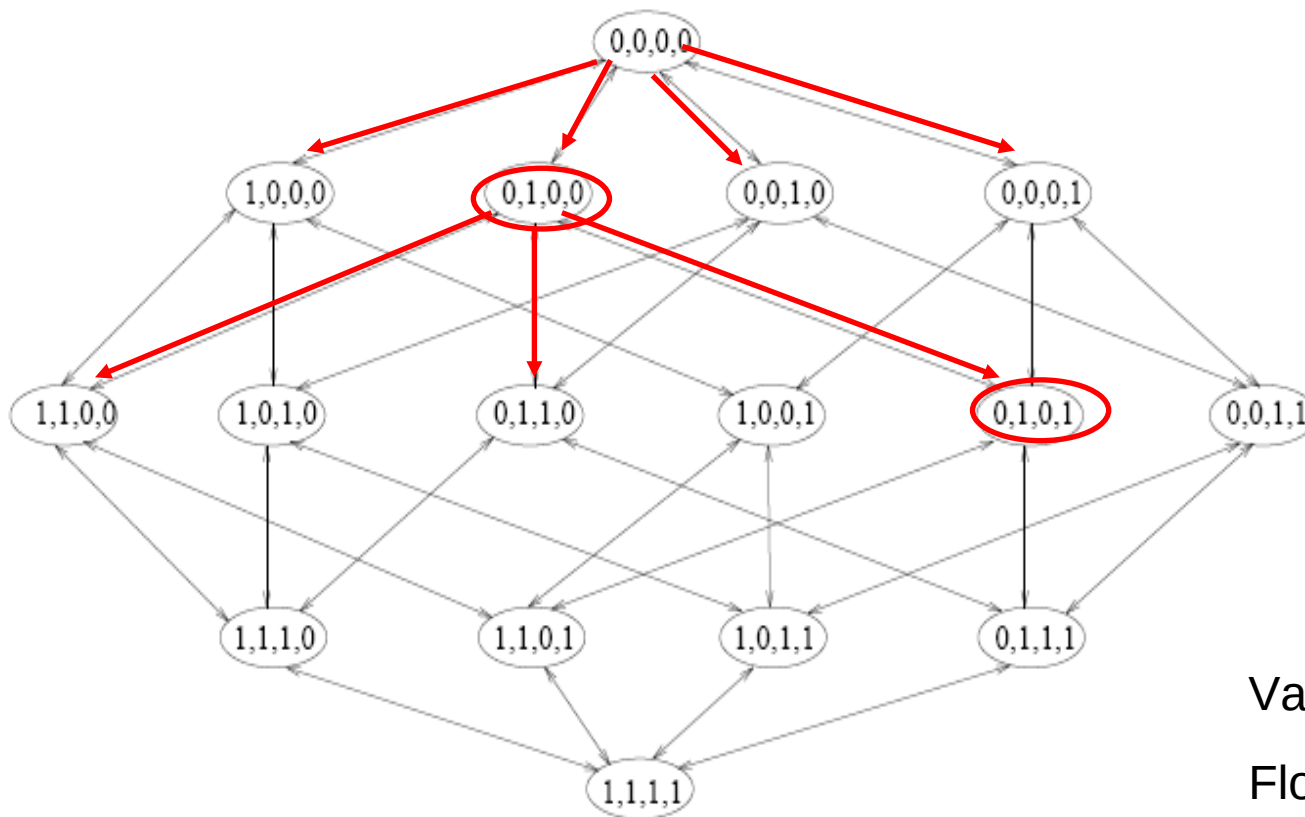
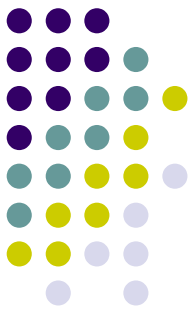
# Wrappers. Forward search



Variantes:

Floating search. 1  
paso adelante, 1 atrás

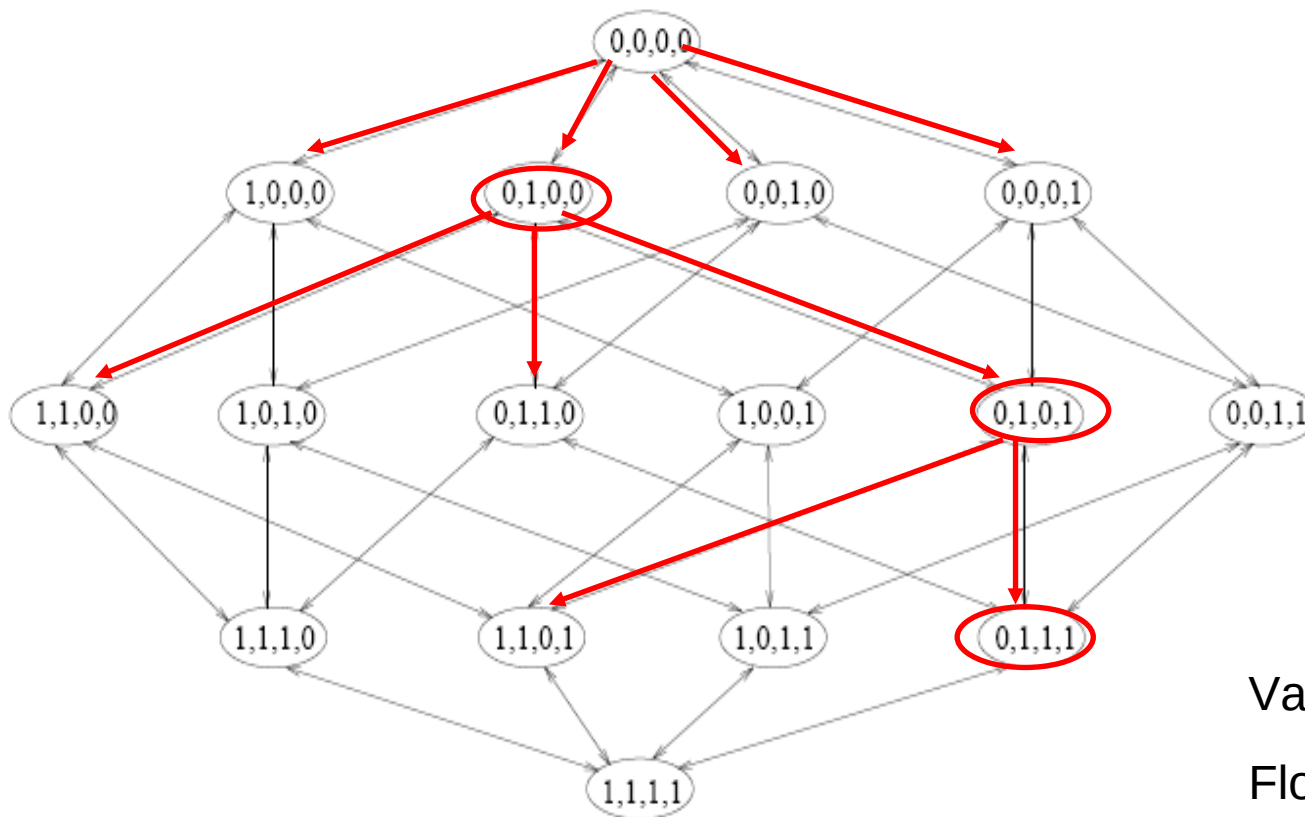
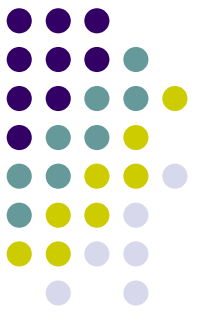
# Wrappers. Forward search



Variantes:

Floating search. 1  
paso adelante, 1 atrás

# Wrappers. Forward search

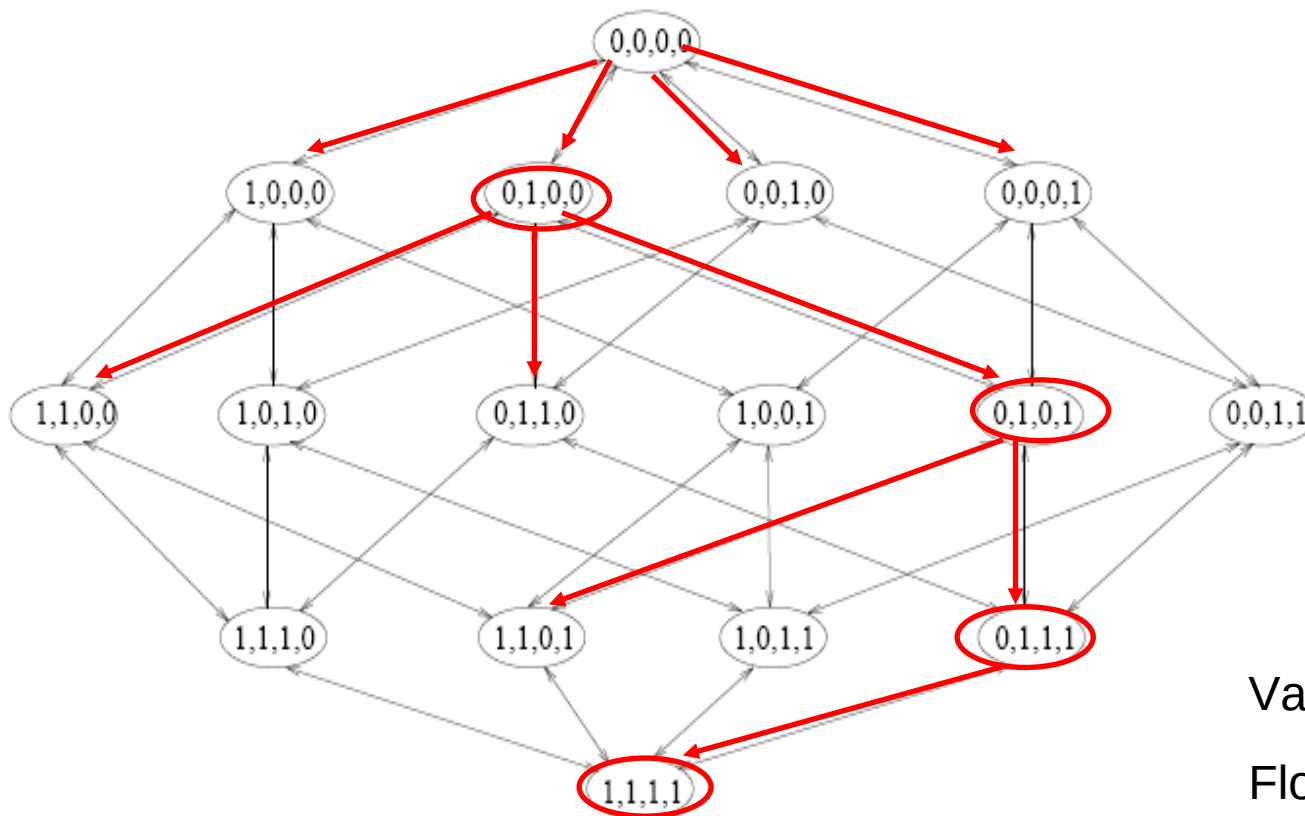
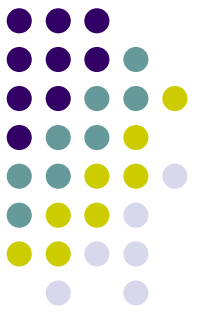


Variantes:

Floating search. 1  
paso adelante, 1 atrás



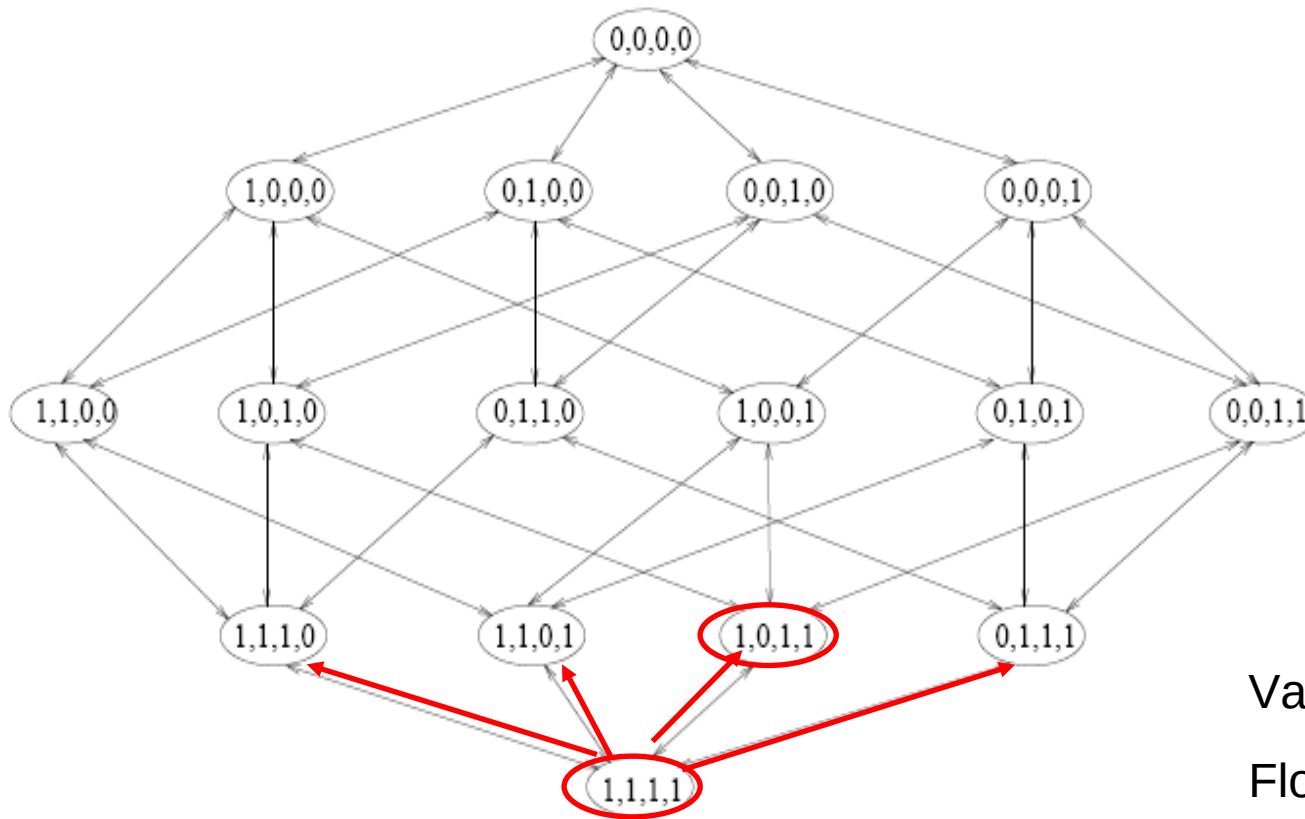
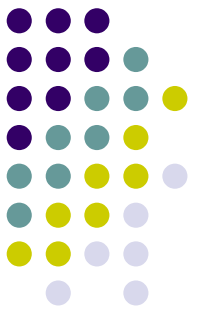
# Wrappers. Forward search



Variantes:

Floating search. 1  
paso adelante, 1 atrás

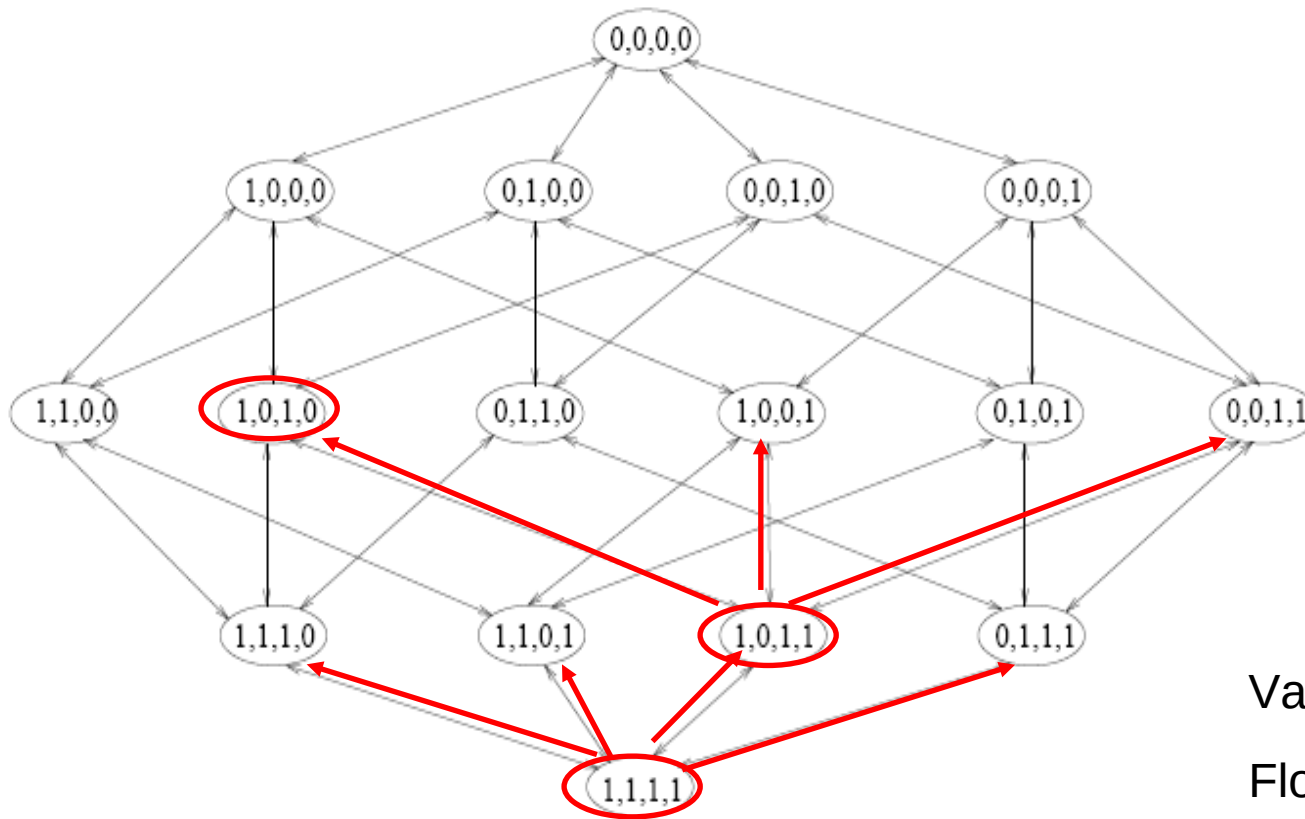
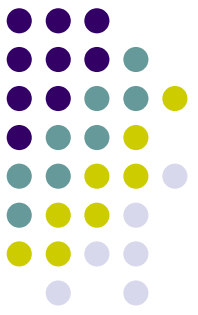
# Wrappers. Backward search



Variantes:

Floating search

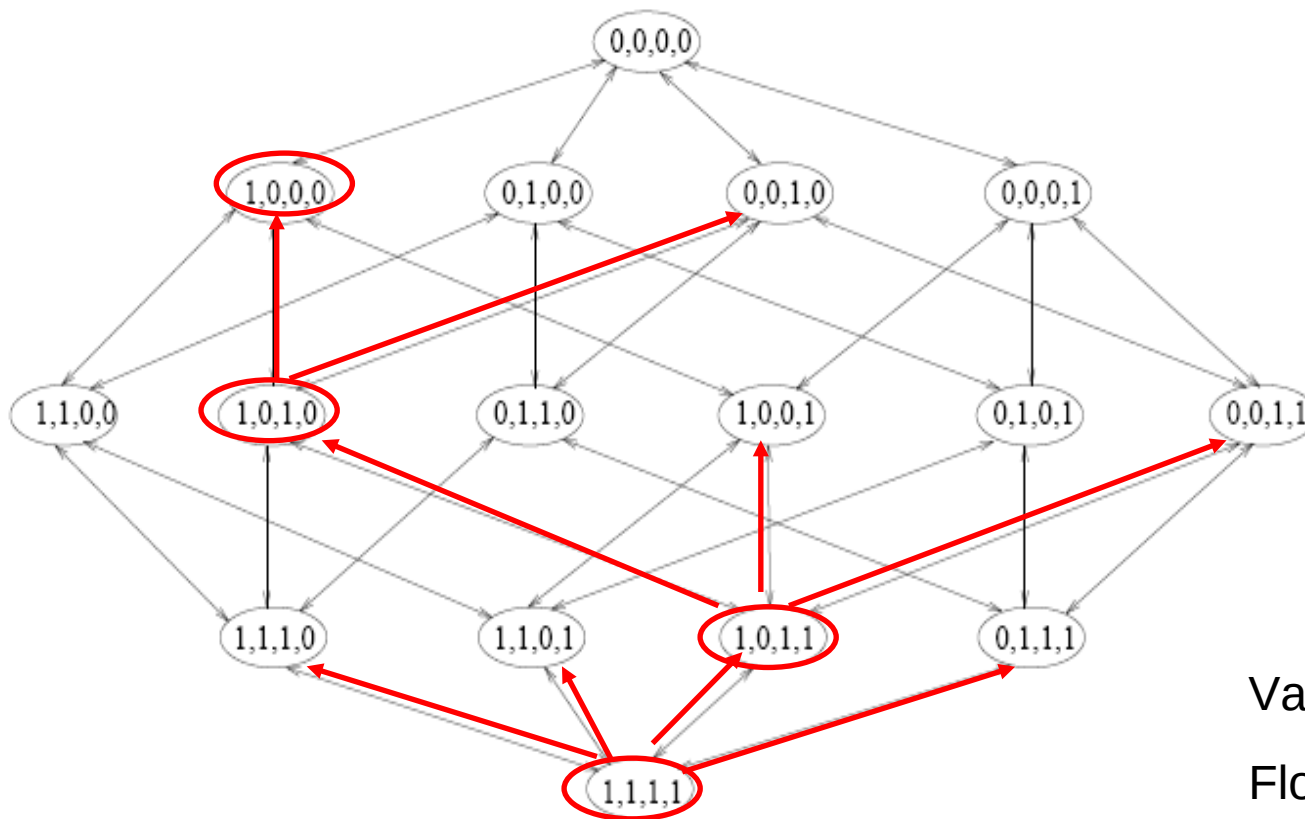
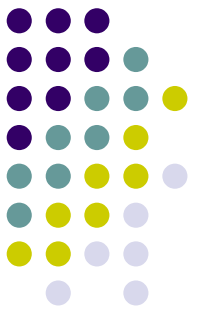
# Wrappers. Backward search



Variantes:

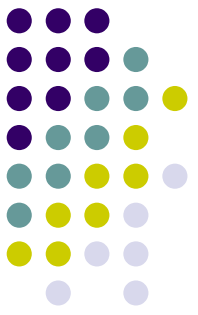
Floating search

# Wrappers. Backward search



Variantes:

Floating search



# Filtros vs. wrappers

- Los dos son heurísticas
- Los filtros:
  - No resuelven el problema de modelado directamente.
  - Suelen tener problemas con variables “conjuntas”
  - Son muy rápidos
- Los wrappers
  - Dan mejores selecciones
  - Son muy pesados
  - Suelen hacer overfitting