Principales Técnicas

Abril de 2016

Contenido

Principales Técnicas

- Las técnicas más relevantes se pueden clasificar según el tipo de problema que se estudia.
- Análisis supervisado:
  1. Clasificación: logístico, árboles, *boosting* de árboles, redes neuronales.
  2. Regresión: regresión lineal, *shrinkage*, redes neuronales, kernels.
- Análisis no supervisado:
  1. Conglomerados (clusters).
  2. Reglas de association.
  3. Análisis digital.

Principales Técnicas: Logístico

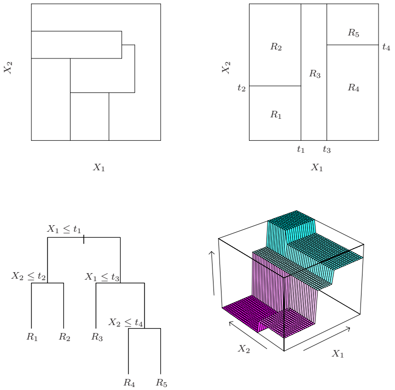
- Supongamos $y \in \{0, 1\}$. El modelo de clasificación logístico supone que:

$$P(y = 1 \mid x) = f(x^T\beta) = \frac{1}{1 + e^{-x^T\beta}} \tag{1}$$

  donde $P(y = 1 | x)$ es la probabilidad de que el resultado sea 1 cuando las características observadas son x.
- La forma de estimar β es mediante el método de máxima versosimilitud. La idea es maximizar:

$$f(x_1^T\beta)^{y_1}(1 - f(x_1^T\beta))^{1-y_1} \times ...f(x_n^T\beta)^{y_n}(1 - f(x_n)^T\beta)^{1-y_n} \tag{2}$$

- La idea es hacer separaciones secuenciales de los datos de acuerdo a cada variable independiente.



- La idea es hacer separaciones secuenciales de los datos de acuerdo a cada variable independiente.
- Sean $R_1, ..., R_M$ un conjunto de regiones.
- Cada región la clasificamos como cero o uno dependiendo de la proporcion de datos observados con resultado cero o uno.
- Sea $p(m, 1)$ La cantidad de unos en la región m como proporción de la cantidad de observaciones en esa región.
- Sea $k(m) = 1$ si $p(m, 1) > p(m, 0)$.
- Definimos el error de clasificación como $L(m) = 1 - p(m, k(m))$
- Ahora repetimos el siguiente algoritmo.

1. Para determinar el primer nodo, la variable que hace separación es aquella que minimiza el error de clasificación.
2. En cada nodo se elige la variable (entre las que no se han usado para separar anteriormente) que minimiza el error de clasificación.
3. Se continua hasta completar M nodos.

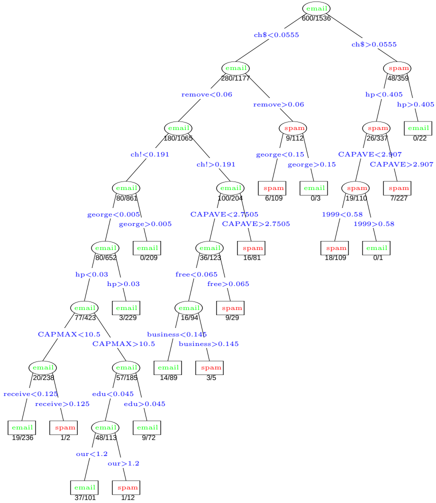
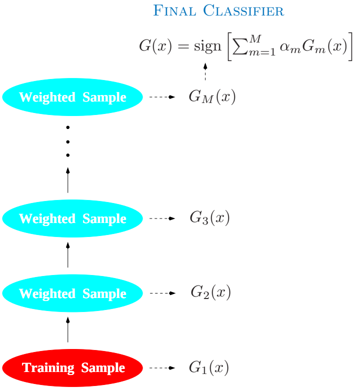


FIGURE 9.5. *The pruned tree for the* spam *example. The split variables are shown in blue on the branches, and the classification is shown in every node. The numbers under the terminal nodes indicate misclassification rates on the test data.*

- Un excelente clasificador cuando se combina con *boosting*



Principales Técnicas: Regresión lineal, shrinkage, *redes neuronales*, kernels

- La característica principal es que permite introducir no linealidades.

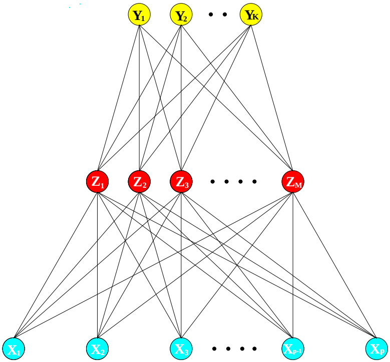


FIGURE 11.2. *Schematic of a single hidden layer, feed-forward neural network.*

- El problema es describir la distribución las variables independientes.
- Permite detectar asociaciones entre variables cuando estas son poco frecuentes.
- Los conceptos claves son:

1. El soporte de un conjunto de variables es la probabilidad de encontrar esos valores entre los valores de las variables independientes.
2. Una regla $A \Rightarrow B$ se le asigna una confianza igual a la probabilidad condicional de B dado A.

- Típicamente se busca un soporte bajo y una confianza alta.

Contenido

Validación y Selección de Modelos

- Objetivos complementarios en un probema de minería de datos.

1. Selección de modelos.
2. Validación de modelos.

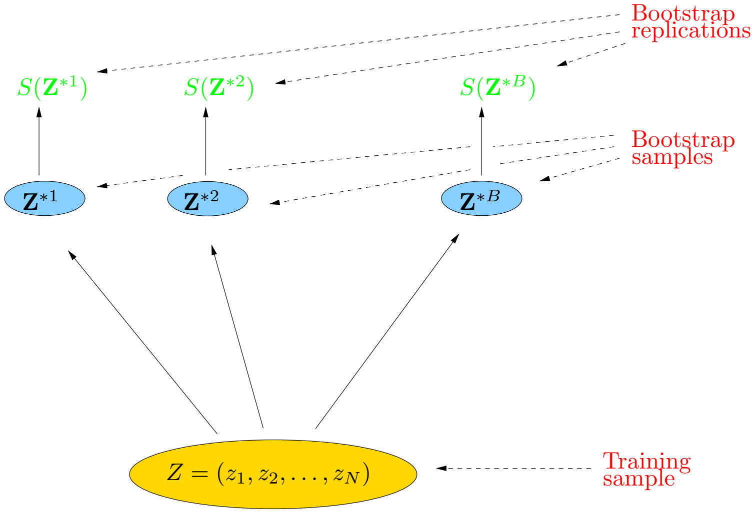
Selección de Modelos

- En situaciones con abundancia de datos una buena práctica es dividir aleatoriamente los datos observados (cuando el número de datos es moderado se puede elegir únicamente entrenamiento y prueba).

- Para datos panel: Tratar los datos como un pool si la dimensión temporal no es tan relevante y validar fuera de muestra (modelos de scoring, deserción).
- Para datos de series de tiempo: procedimiento estandar de series de tiempo o emular una muestra de entrenamiento y prueba que cubra el msmo horizonte.

Bootstrap

- Cuando no hay muchos datos una alternativa es hacer bootstrapping.



Bootstrap

- No es tan útil para hacer predicciones fuera de muestra (intuitivamente, los mismos datos con los que se entrena se usan para validar).
- Para clases desbalanceadas una modificación de bootstapping es muy útil para mejorar el poder de clasificación: sub Bagging.
- La técnica de validación cruzada mitiga este problema.

Validación cruzada

1	2	3	4	5
Train	Train	Validation	Train	Train

Validación cruzada y regularización

- Una forma usual de elegir un parámetro de regularización (i.e., penalización) es hacer validación cruzada, índice H.

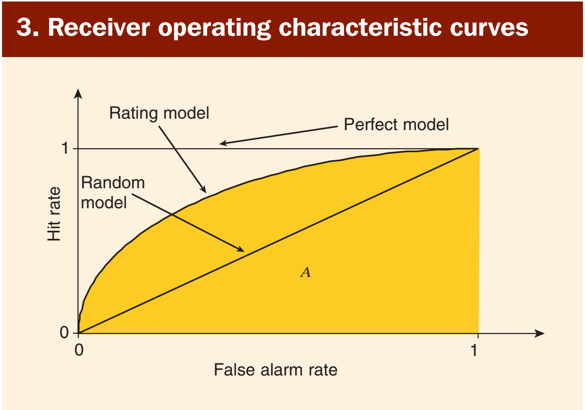
Validación de Modelos

- Modelos de regresión: AIC, R^2, MAPE, etc.
- Modelo de clasificación: Curva ROC y curva de calibración

Curva ROC

- La curva ROC y el área bajo la curva es uno de los métodos más importantes de validación para problemas de clasificación.

Curva ROC



- Los modelos de clasificación binaria se pueden extender a clasificación en múltiples categorias.

Curva ROC: Explicación

- Sea M el total de unos, M el total de ceros y $T = M + B$.
- El verdaderos positivos (hit rate) se define como: Los que resultaron malos entre los que se pronosticaron malos sobre M.
- Los falsos positivos se define como: Los que salieron buenos entre los que se pronosticaron como malos sobre B
- Entre más cerca este la curva del eje vertical y el eje horizontal superior mmejor el poder de clasificación global: para una misma tasa de falsos positivos, una mayor tasa de acierto.
- Obsérvese que ambos modelos pronostican diferentes números de malos.

Curva ROC: Explicación

- El modelo aleatorio tiene una curva ROC de 45 grados.

Curva de calibración

- Mide el error entre las frecuencias pronosticadas de un evento y las frecuencias observadas de ocurrencia.
- En las aplicaciones de aprendizaje de máquinas a medicina es estándar y se usa una prueba χ^2 para determinar la significacia estadística de la diferencia.
- Es manipulable.