Redes Bayesianas: Clasificación supervisada (I)

Aritz Pérez[1] Borja Calvo[2]

Basque Center for Applied Mathematics

UPV/EHU

Donostia, Febrero de 2015

Bibliografía

- K.P. Murphy (2012). Machine Learning: **A Probabilistic Perspective**. The MIT Press.

Clasificación supervisada

- Tenemos un conjunto de variables **predictoras** $\mathbf{X}$, una variable **clase** C y un conjunto de **entrenamiento** completo $\mathcal{D} = \{(\mathbf{x}^1, c^1), ..., (\mathbf{x}^N, c^N)\}$
- **Clasificación** supervisada: construir un mapeo del espacio $\Omega_{\mathbf{X}}$ en Ω_C.
- **Clasificador**: f es una **función** que asigna un valor c a cada caso no etiquetado $\mathbf{x}$.
- **Objetivo**: Construir un **clasificador** f que minimice el **error** de clasificación a partir de los datos
- Algoritmo de aprendizaje de clasificadores A.

Extensiones

- Múltiples variables clase: **multi-dimensional**
- Conjunto de entrenamiento incompleto: **semi-supervisada**
- Conjunto de entrenamiento sin clase: **no supervisada**
- ...

Aproximación generativa

Regla de Bayes

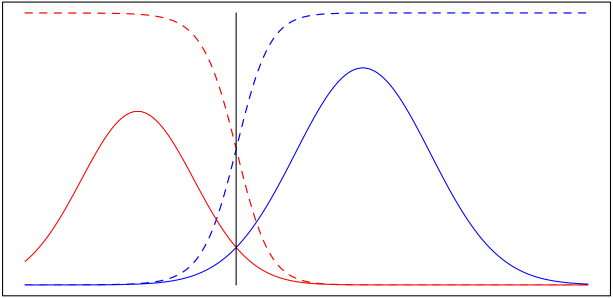
$$
\begin{aligned}
c^* &= \arg\max_c p_M(c|\mathbf{x}) \\
&= \arg\max_c p_M(\mathbf{x}, c)
\end{aligned}
$$

- Construir **modelo** probabilista, p_M
- Clasificar nuevos casos en la **clase más probable**
- **Clasificador de Bayes**, $\arg\max_c p(c|\mathbf{x})$: menor error de clasificación
- Construir **red Bayesiana** $M = (G, \mathbf{\Theta})$ y emplear la **regla de Bayes** $f_M(x) \equiv \arg\max_x p_M(\mathbf{x}, c)$

Otras aproximaciones

- Aproximación **condicional**: basada un modelo de la probabilidad condicionada, $p_M(c|\mathbf{x})$, y la regla de Bayes
- Aproximación **discriminativa**: construir una función f directamente

Aproximaciones



Ventajas de la aproximación generativa

- Permite obtener la **probabilidad condicionada** $p(C|\mathbf{x})$
- Permite generar nuevos casos: **muestreo**
- **Combinación de modelos** parciales de forma probabilista: $p(\mathbf{x}_A|c)p(\mathbf{x}_B|c)p(c)$
- Incorporar **información a priori**: estadística Bayesiana
- Reutilizar el **aprendizaje** de redes Bayesianas

Ventajas de la aproximación generativa

Permite solucionar **diferentes problemas** con el mismo modelo:

- Clasificación empleando casos con **valores perdidos**
- Detección de casos improbables, $p(\mathbf{x})$: **outlier detection**
- Clasificación empleando matriz de **costes**
- Emplear una región de **rechazo**
- Clases **desbalanceadas**: $p(c) >> p(c')$

Inconvenientes de la aproximación generativa

- Aproxima un **problema más difícil** para resolver el de clasificación
- Requiere **más parámetros**
- Modela **información irrelevante** para la clasificación, $p(\mathbf{x})$

Error de clasificación

Definición

Probabilidad de cometer un error:

$$\epsilon(f) = \sum_{(\mathbf{x},c)} p(\mathbf{x}, c)\delta(c \neq f(\mathbf{x}))$$

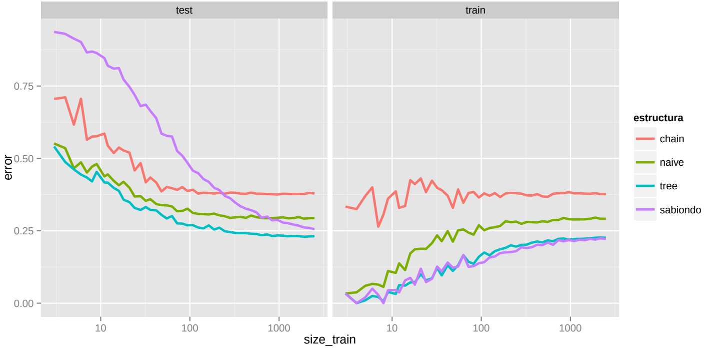
- Construir clasificadores que **minimizen el error**
- Valor **desconocido**: $p(\mathbf{x}, c)$?
- Necesita ser **aproximado/estimado** empleando un conjunto de **test**:

$$\hat{\epsilon}(f; \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{(x,c)\in\mathcal{T}} 1(c \neq f(\mathbf{x}))$$

Paralelismo con generalizacion y ajuste

- **Error** en el conjunto de **entrenamiento** $\simeq$ **ajuste**
- Error en el conjunto de **test** $\simeq$ **generalización**
- Evitar **sobreajuste**
- **Equilibrio** entre número de **parámetros** y de **casos** del conjunto de entrenamiento

Paralelismo con generalizacion y ajuste



Estimación del error

Principios reñidos

1. No emplear en el conjunto de test los casos del entrenamiento: **optimismo**
2. No emplear menos casos de los disponibles para entrenar: **pesimismo**

Principales estimadores del error

- Reemplazamiento (incumple 1)
- Holdout (incumple 2)
- Validación cruzada (incumple 2)
- Bootstrap (incumple 2)

Siempre que se pueda...

- **Repetición**: Reduce la varianza
- **Estratificación**: Reduce el sesgo y la varianza