Modelos lineales para Clasificación

Hernán Felipe García Arias M.Sc.

Aprendizaje de Máquina - UTP

2018

Contenido

Definiciones (I)

- **Objetivo.** Tomar un vector de entrada, $\mathbf{x}$, y asignarlo a una de K clases $\mathcal{C}_k$, para $k = 1, \ldots, K$.

- **Espacio de entrada.** Se divide en regiones de decisión $\mathcal{R}_k$.

- **Líneas o superficies de decisión.** Separan las regiones de decisión.

Definiciones (II)

- **Modelo lineal.** Las superficies de decisión son funciones lineales de $\mathbf{x}$.

- Las superficies están definidas por hiperplanos de dimensión $D - 1$, en un espacio de D dimensiones.

- Si los datos se pueden separar linealmente por una superficie de decisión lineal, se dice que los datos son *linealmente separables*.

Definiciones (III)

- En regresión, t_n representaba un número real asociado a la entrada $\mathbf{x}_n$.

- En clasificación, $\mathbf{t}_n$ representa un vector, en codificación 1 de K.

- Tres enfoques al problema de clasificación
  - Funciones discriminantes.
  - Modelos generativos, $p(\mathbf{x}, \mathcal{C}_k)$.
  - Modelos discriminativos, $p(\mathcal{C}_k|\mathbf{x})$.

Contenido

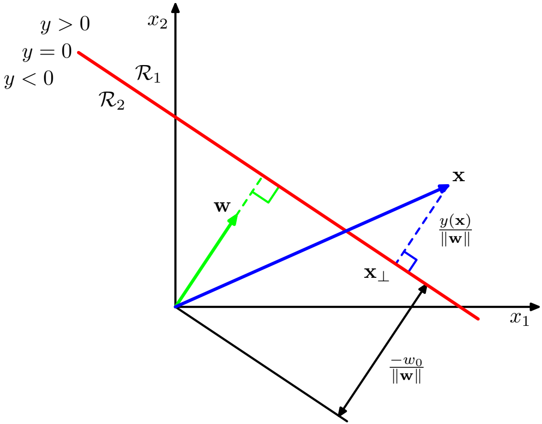
Función discriminante (I)

- Función $y(\mathbf{x}) : \mathbf{x} \to k$, $k \in \{1, \ldots, K\}$.
- Dos clases: $y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$.
- $\mathbf{w}$ es el vector de pesos, y w_0 es la tendencia.
- $\mathbf{x} \in \mathcal{C}_1$, si $y(\mathbf{x}) > 0$. De lo contrario, $\mathbf{x} \in \mathcal{C}_2$.
- La línea de decisión o superficie de decisión es $y(\mathbf{x}) = 0$.
- El vector $\mathbf{w}$ es ortogonal a la superficie de decisión.

Función discriminante (II)

Distancia de $y(\mathbf{x})$ al origen: $-w_0/\|\mathbf{w}\|$.

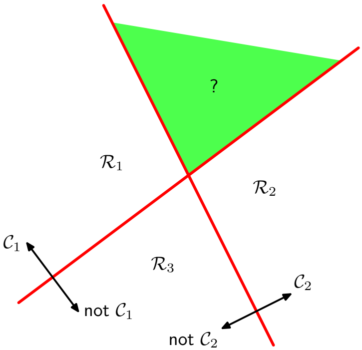
Distancia de un punto $\mathbf{x}$ a $y(\mathbf{x})$: $y(\mathbf{x})/\|\mathbf{w}\|$.



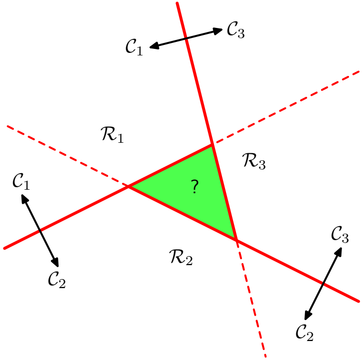
Se puede hace $\widetilde{\mathbf{x}} = (x_0, \mathbf{x})$, y $\widetilde{\mathbf{w}} = (w_0, \mathbf{w})$, $y(\mathbf{x}) = \widetilde{\mathbf{w}}^\top \widetilde{\mathbf{x}}$.

Múltiples clases (I)

Construir un clasificador de K clases a partir de clasificadores de 2 clases.



1 vs. todos



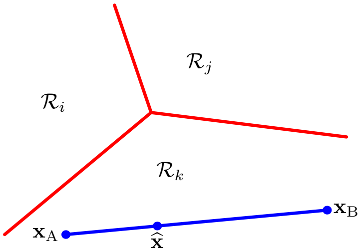
1 vs 1

Múltiples clases (II)

- **Solución:** discriminante de K clases con K funciones lineales

$$y_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0}.$$

- $\mathbf{x} \in C_k$, si $y_k(\mathbf{x}) > y_j(\mathbf{x})$, $k \neq j$.
- Regiones de decisión conectadas de manera simple y conexas.

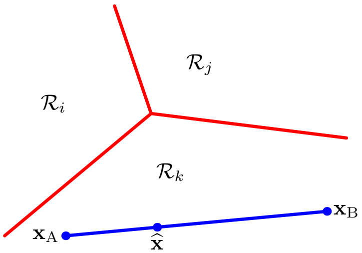


Múltiples clases (III)

- Sean $\mathbf{x}_A, \mathbf{x}_B \in \mathcal{R}_k$.
- Si $\widehat{\mathbf{x}}$ es un punto sobre la línea que une a $\mathbf{x}_A$ y $\mathbf{x}_B$,

$$\widehat{\mathbf{x}} = \lambda \mathbf{x}_A + (1-\lambda)\mathbf{x}_B,$$
$$y_k(\widehat{\mathbf{x}}) = \lambda y_k(\mathbf{x}_A) + (1-\lambda) y_k(\mathbf{x}_B).$$

- Como $y_k(\mathbf{x}_A) > y_j(\mathbf{x}_A)$, y $y_k(\mathbf{x}_B) > y_j(\mathbf{x}_B)$, $(k \neq j)$ luego $y_k(\widehat{\mathbf{x}}) > y_j(\widehat{\mathbf{x}})$.



Contenido

Mínimos cuadrados (I)

- Cada clase $\mathcal{C}_k$ se describe por su propio modelo lineal

$$y_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0},$$

para $k \in \{1, \ldots, K\}$.

- Agrupando,

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^\top \widetilde{\mathbf{x}},$$

donde $\widetilde{\mathbf{W}}$ tiene columnas $\widetilde{\mathbf{w}}_k = (w_{k0}, \mathbf{w}_k)^\top$, y $\widetilde{\mathbf{x}} = (1, \mathbf{x}^\top)^\top$.

Mínimos cuadrados (II)

- Sea un conjunto de entrenamiento $\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$.
- La matrix $\mathbf{T}$ tiene vectores fila $\mathbf{t}_n^\top$, y la matriz $\widetilde{\mathbf{X}}$ vectores fila $\widetilde{\mathbf{x}}_n$.
- La función de error cuadrático está dada como

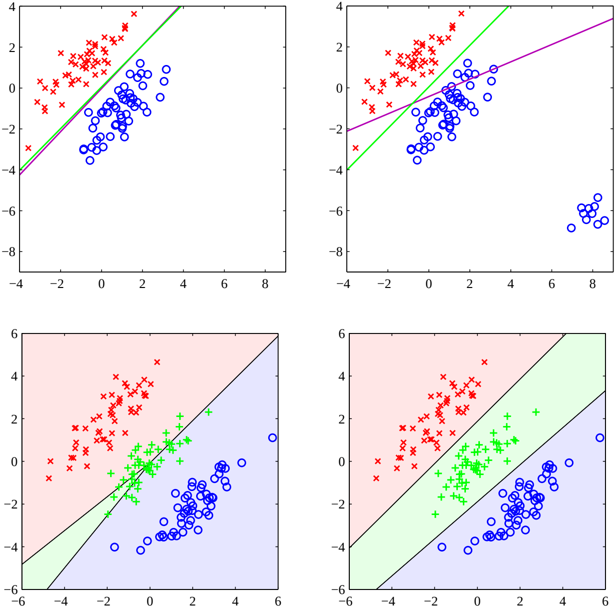
$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2} \operatorname{tr} \left\{ (\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^\top (\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T}) \right\}.$$

- Minimizando e igualando a cero se obtiene

$$\widetilde{\mathbf{W}}_{MSE} = \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} \right)^{-1} \widetilde{\mathbf{X}}^\top \mathbf{T} = \widetilde{\mathbf{X}}^\dagger \mathbf{T},$$

  donde $\widetilde{\mathbf{X}}^\dagger$ es la seudo inversa de $\widetilde{\mathbf{X}}$.
- La función discriminante está dada por $\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}_{MSE}^\top \widetilde{\mathbf{x}} = \mathbf{T}^\top \left(\widetilde{\mathbf{X}}^\dagger \right)^\top \widetilde{\mathbf{x}}$.

Mínimos cuadrados: inconvenientes



Análisis Discriminante de Fisher (I)

- La idea es proyectar los datos a un espacio de menor dimensionalidad donde la clasificación sea más sencilla.
- Sea $\mathbf{x} \in \mathbb{R}^D$.
- Se proyecta a una dimensión usando

$$y = \mathbf{w}^\top \mathbf{x}.$$

- Se establece un umbral y_0, y se clasifica un nuevo punto como de la clase $\mathcal{C}_1$ si $y \geq y_0$, o de la clase $\mathcal{C}_2$, si pasa lo contrario.
- La idea es escoger $\mathbf{w}$ de manera que maximice la separabilidad de las clases.

Análisis Discriminante de Fisher (II)

- Sea un poblema biclase, con vectores de media

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n, \qquad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n.$$

- Una medida de la separación entre las clases es

$$m_1 - m_2 = \mathbf{w}^\top (\mathbf{m}_1 - \mathbf{m}_2).$$

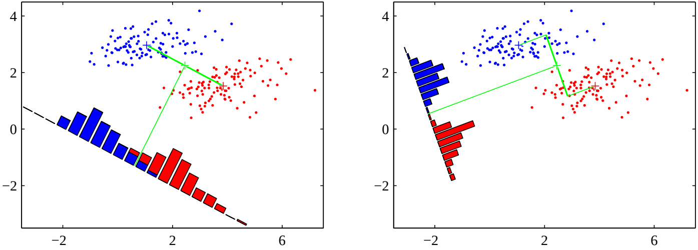
- En la expresión anterior **w** se puede hacer muy grande, pero se limita para que tenga longitud ordinaria.

Análisis Discriminante de Fisher (III)

- No sólo se quiere minimizar la distancia entre las medias, si no también minimizar la variabilidad de las muestras en cada clase.
- La varianza intraclase se obtiene de los vectores transformados de la clase $\mathcal{C}_k$ como

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2,$$

donde $y_n = \mathbf{w}^\top \mathbf{x}_n$.



Análisis Discriminante de Fisher (IV)

- El criterio de Fisher se define como el radio de la varianza entre clases sobre la varianza intraclase

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}.$$

- Haciendo los cambios apropiados se tiene

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}},$$

  donde

$$\begin{aligned}
\mathbf{S}_B &= (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top, \\
\mathbf{S}_W &= \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top.
\end{aligned}$$

- $\mathbf{S}_B$ es la matriz de covarianza *entre clases*, y $\mathbf{S}_W$ es la matriz de covarianza *intra clases*.

Análisis Discriminante de Fisher (V)

- Derivando $J(\mathbf{w})$ con respecto a $\mathbf{w}$ e igualando a cero se tiene

$$(\mathbf{w}^\top \mathbf{S}_B \mathbf{w})\mathbf{S}_W \mathbf{w} = (\mathbf{w}^\top \mathbf{S}_W \mathbf{w})\mathbf{S}_B \mathbf{w}.$$

- Lo que importa de $\mathbf{w}$ es su dirección, no su magnitud.

- Premultiplicando por $\mathbf{S}_W^{-1}$ se encuentra que

$$\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1).$$

- El resultado se conoce como el *discriminante lineal de Fisher*.

Algoritmo del Perceptrón (I)

- Dos clases. Los datos de entrada se transforman como $\phi(\mathbf{x})$.

- El modelo lineal tiene la forma

$$y(\mathbf{x}) = f(\mathbf{w}^\top \phi(\mathbf{x})),$$

  donde $f(\cdot)$ es la función escalón unitario.

- En el perceptrón se asume $t = +1$, para $\mathcal{C}_1$, y $t = -1$ para $\mathcal{C}_2$.

Algoritmo del Perceptrón (II)

- La función a minimizar se conoce como el criterio del perceptrón.

$$E_P(\mathbf{w}) = -\sum_{n \in \mathcal{M}} \mathbf{w}^\top \phi(\mathbf{x}_n) t_n,$$

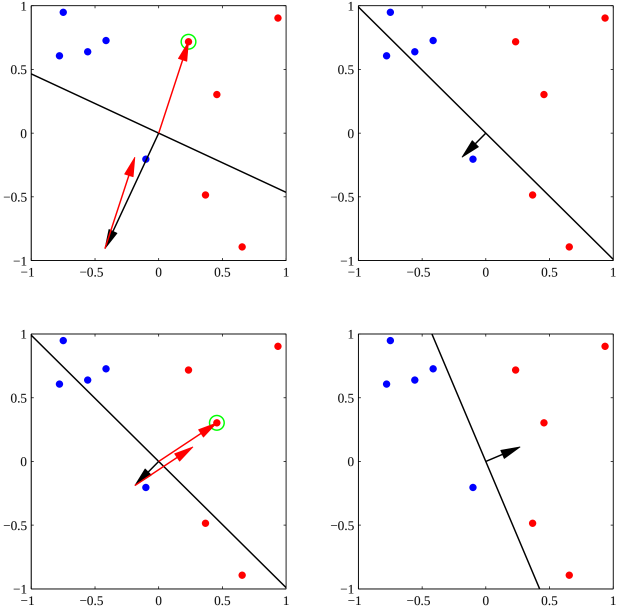
donde $\mathcal{M}$ denota el conjunto de patrones incorrectamente clasificados.

- Aplicando el algoritmo de gradiente descendente estocástico a esta función, se tiene

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_P(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta \phi(\mathbf{x}_n) t_n,$$

donde η se conoce como la razón de aprendizaje, y τ indexa los pasos del algoritmo.

Algoritmo del Perceptrón (III)



Contenido

Introducción

- En los modelos generativos se modelan la densidad de clase condicional $p(\mathbf{x}|\mathcal{C}_k)$, y las funciones de probabilidad a priori, $p(\mathcal{C}_k)$.

- Ambas probabilidades se usan para calcular el posterior $p(\mathcal{C}_k|\mathbf{x})$.

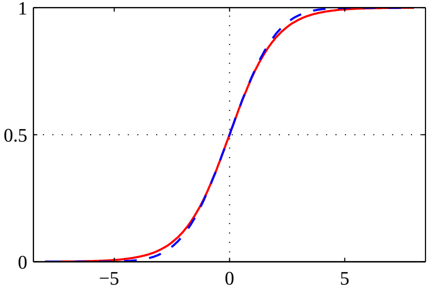
Modelo generativo (I)

Para el caso biclase,

$$
\begin{aligned}
p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1 + \frac{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}} \\
&= \frac{1}{1 + \exp\left\{\ln\left[\frac{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}\right]\right\}} = \frac{1}{1 + \exp\left\{-\ln\left[\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}\right]\right\}} \\
&= \frac{1}{1 + \exp(-a)} = \sigma(a),
\end{aligned}
$$

donde $a = \ln\left[\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}\right]$, y $\sigma(a) = 1/(1 + \exp(-a))$ es la función logística sigmoidal .

Modelo generativo (II)



Para el caso $K > 2$ se tiene

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)},$$

con $a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$. (función soft-max).

Modelo generativo: entradas continuas (I)

- Clase condicional Gaussiana, matrix de covarianza igual para todas las clases,

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}.$$

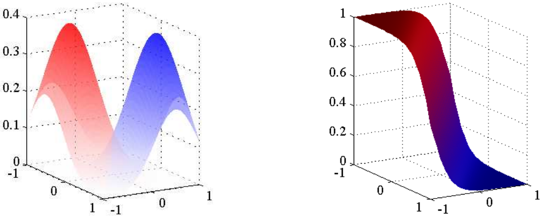
- Considérese dos clases. Para la clase $\mathcal{C}_1$ se tiene

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + w_0),$$

donde

$$\begin{aligned}
\mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\
w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}.
\end{aligned}$$

Modelo generativo: entradas continuas (II)



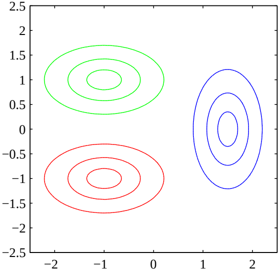
Para el caso $K > 2$ se tiene

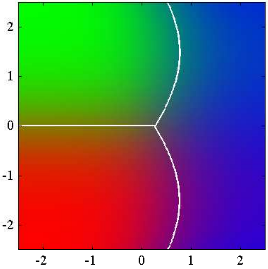
$$a_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0},$$

donde se ha definido $\mathbf{w}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$, $w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k)$.

Modelo generativo: entradas continuas (III)

Si las clases no comparten la misma matriz de covarianza, las regiones de decisión son cuadráticas





Máxima verosimilitud (I)

- Los valores de $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$, $p(\mathcal{C}_1)$ y $p(\mathcal{C}_2)$ se determinan usando máxima verosimilitud.
- Sean dos clases y un conjunto de datos $\{\mathbf{x}_n, t_n\}_{n=1}^{N}$.
- $t_n = 1$ denota $\mathcal{C}_1$, y $t_n = 0$ denota $\mathcal{C}_2$.
- Las probabilidades a priori se escriben como $p(\mathcal{C}_1) = \pi$ y $p(\mathcal{C}_2) = 1 - \pi$.

Máxima verosimilitud (II)

- Para un punto $\mathbf{x}_n$ de la clase $\mathcal{C}_1$ se tiene $t_n = 1$ y así

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathbf{x}_n|\mathcal{C}_1)p(\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}).$$

- De forma similar, para la clase $\mathcal{C}_2$

$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathbf{x}_n|\mathcal{C}_2)p(\mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}).$$

- Luego,

$$p(t_n|\pi, \boldsymbol{\mu}_1\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \begin{cases} p(\mathbf{x}_n, \mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), & \text{si } t_n = 1 \\ p(\mathbf{x}_n, \mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), & \text{si } t_n = 0. \end{cases}$$

- Lo anterior se puede escribir de forma resumida como

$$p(t_n|\pi, \boldsymbol{\mu}_1\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = [\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n}[(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

Máxima verosimilitud (III)

- Asumiendo que los datos son iid,

$$p(\mathbf{t}|\pi, \boldsymbol{\mu}_1 \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1 - \pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}.$$

- Se realiza la maximización de $\ln p(\mathbf{t}|\pi, \boldsymbol{\mu}_1 \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, el logaritmo de la verosimilitud, con respecto a $\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$,

Máxima verosimilitud: solución para π, $\boldsymbol{\mu}_1$, y $\boldsymbol{\mu}_2$.

- Se puede demostrar que

$$\pi_{ML} = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2},$$

donde N_1 denota el número de puntos de la clase $\mathcal{C}_1$, y N_2 el número de puntos de la clase $\mathcal{C}_2$.

- Igualmente, se puede demostrar que

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n, \qquad \boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1 - t_n)\mathbf{x}_n.$$

Máxima verosimilitud: solución para Σ.

Finalmente, para $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \mathbf{S} = \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2$$

$$\mathbf{S}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^\top,$$

$$\mathbf{S}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^\top.$$

Características discretas

- Sea x_i una variable que toma valores binarios $\{0, 1\}$.
- Se asume que los x_i son independientes condicionados a la clase $\mathcal{C}_k$.
- Las distribuciones de clase condicional se obtienen como

$$p(\mathbf{x}|\mathcal{C}_k) = \prod_{i=1}^{D} \mu_{i,k}^{x_i}(1 - \mu_{i,k})^{1-x_i}.$$

  donde $\mu_{i,k}$ es la probabilidad $p(x_i = 1|\mathcal{C}_k)$.

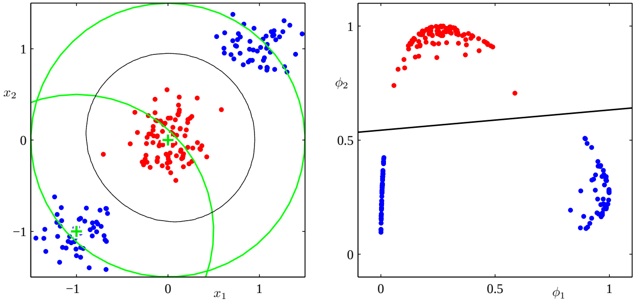
- Sustituyendo en la ecuación $a_k(\mathbf{x}) = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$,

$$a_k(\mathbf{x}) = \sum_{i=1}^{D}\{x_i \ln \mu_{i,k} + (1 - x_i)\ln(1 - \mu_{i,k})\} + \ln p(\mathcal{C}_k).$$

Contenido

Introducción

- Se modela directamente la función de probabilidad a posteriori $p(\mathcal{C}_k|\mathbf{x})$.
- En general, se necesita determinar un número menor de parámetros.
- Se pueden introducir funciones base $\phi(\mathbf{x})$. En el espacio transformado la separación podría ser lineal.



Regresión Logística (I)

- En forma general

$$p(\mathcal{C}_1|\phi) = y(\phi) = \sigma(\mathbf{w}^\top \phi),$$

  donde $\sigma(\cdot)$ es la función logística sigmoidal.

- En estadística este modelo se conoce como *regresión logística*.

- Sea un conjunto de datos $\{\phi_n, t_n\}_{n=1}^N$, con $\phi_n = \phi(\mathbf{x}_n)$ y $t_n \in \{0, 1\}$.

- La función de verosimilitud se define como

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n},$$

  donde $\mathbf{t} = [t_1 \cdots t_N]^\top$, y $y_n = p(\mathcal{C}_1|\phi_n)$.

Regresión Logística (II)

- Se define una función de error tomando el logaritmo negativo de la función de verosimilitud

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} t_n \ln y_n + (1 - t_n) \ln(1 - y_n).$$

donde $\sigma(\cdot)$ es la función logística sigmoidal.

- El gradiente de la función de error con respecto a **w** sigue la forma

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n = \mathbf{\Phi}^{\top}(\mathbf{y} - \mathbf{t}),$$

donde $\mathbf{y} = [y_1 \cdots y_N]^{\top}$.

Mínimos cuadrados reponderados iterativos (I)

- La función de error puede minimizarse usando el algoritmo de *Newton-Raphson*, que toma la forma

$$\mathbf{w}^{(\text{nuevo})} = \mathbf{w}^{(\text{viejo})} - \mathbf{H}^{-1}\nabla E(\mathbf{w}),$$

  donde $\mathbf{H}$ es la matriz Hessiana, $\mathbf{H} = \nabla\nabla E(\mathbf{w})$.

- La matriz Hessiana se calcula como

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1 - y_n)\phi_n\phi_n^\top = \mathbf{\Phi}^\top\mathbf{R}\mathbf{\Phi},$$

  donde $\mathbf{R}$ es una matriz diagonal de $N \times N$ con elementos $R_{nn} = y_n(1 - y_n)$.

Mínimos cuadrados reponderados iterativos (II)

- La solución para **w** debe encontrarse de forma iterativa, debido a que los elementos de **R** dependen de **w**, a través de y_n.

- La solución para **w** se puede escribir como

$$\mathbf{w}^{(\text{nuevo})} = \left(\boldsymbol{\Phi}^\top \mathbf{R} \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \mathbf{R}\mathbf{z},$$

  donde $\mathbf{z} = \boldsymbol{\Phi}\mathbf{w}^{(\text{viejo})} - \mathbf{R}^{-1}(\mathbf{y} - \mathbf{t})$.

- La solución anterior es parecida a la solución de mínimos cuadrados para el problema de regresión lineal.

- Las ecuaciones normales se deben aplicar iterativamente.

- Por esta razón este algoritmo se conoce como *mínimos cuadrados reponderados iterativos* (IRLS - Iterative Reweighted Least Squares).