APA: Aprenentatge Automàtic (TEMES 5 i 6)

Grau en Enginyeria Informàtica - UPC (2017/18)

Lluís A. Belanche, belanche@cs.upc.edu

Entrega: 11 Desembre 2017

Els problemes marcats **[G]** són de grup; els problemes/apartats marcats **[R]** són per fer-se en R

Objectius:

1. Conéixer i saber derivar la funció d'error més adequada per classificació
2. Saber crear i aplicar classificadors discriminatius Bayesians
3. Entendre i aplicar la regressió logística

Problema 1

Considerem un problema de classificació en dues classes, en les quals es disposa de les probabilitats de cada classe $P(C_1)$ i $P(C_2)$. Considerem tres possibles regles per classificar un objecte:

1. (R_1) Predir la classe més probable
2. (R_2) Predir la classe C_1 amb probabilitat $P(C_1)$
3. (R_3) Predir la classe C_1 amb probabilitat 0.5

Es demana:

1. Donar les probabilitats d'error $P_i(\text{error})$ de les tres regles, $i = 1, 2, 3$
2. Demostrar que $P_1(\text{error}) \leq P_2(\text{error}) \leq P_3(\text{error})$

. . . . . .

Problema 2 Funcions d'error per classificació [G]

L'objectiu dels models probabilístics *discriminatius* per classificació és modelar les probabilitats a posteriori $P(C_k|\boldsymbol{x})$ per a cada classe k. En tasques de classificació binària (dues classes, C_1 i C_2), modelem amb una funció $y(\boldsymbol{x}) = P(C_1|\boldsymbol{x})$; llavors $1 - y(\boldsymbol{x}) = P(C_2|\boldsymbol{x})$. Tenim una mostra aleatòria simple D de llargada N del mecanisme $p(t, \boldsymbol{x})$, que escrivim $D = \{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$, on $\boldsymbol{x}_n \in \mathbb{R}^d$ i $t_n \in \{0, 1\}$. Prenem la convenció que $t_n = 1$ indica $\boldsymbol{x}_n \in C_1$ i $t_n = 0$ indica $\boldsymbol{x}_n \in C_2$, i modelem:

$$P(t|\boldsymbol{x}) = \begin{cases} y(\boldsymbol{x}) & \text{si } \boldsymbol{x}_n \in C_1 \\ 1 - y(\boldsymbol{x}) & \text{si } \boldsymbol{x}_n \in C_2 \end{cases}$$

que pot ser més convenientment expressat com $P(t|\boldsymbol{x}) = y(\boldsymbol{x})^t(1 - y(\boldsymbol{x}))^{1-t}$, $t = 0, 1$. Aquesta és una distribució de Bernoulli, la qual cosa permet d'obtenir una funció d'error amb criteris ben fonamentats.

1. Construïu la funció log-versemblança de la mostra i proposeu una funció d'error a partir d'ella.
2. Generalitzeu el resultat a un número arbitrari $K \geq 2$ de classes.

. . . . . .

Problema 3 Model probabilístic generatiu per variables binàries

Considerem el cas de tenir d variables binàries $x_i \in \{0,1\}$ en un problema de classificació en K classes, $C_1, \ldots, C_K$. La distribució conjunta $P(\boldsymbol{x}) = P(x_1, \ldots, x_d)$ requereix en principi el coneixement de $2^d - 1$ números (les respectives probabilitats de cada combinació) per cada classe, la qual cosa no és factible.

Decidim doncs treballar amb distribucions condicionals (per cada classe k) de la forma:

$$P(\boldsymbol{x}|C_k) = \prod_{i=1}^{d} p_{ki}^{x_i}(1-p_{ki})^{1-x_i}$$

on p_{ki} és la probabilitat de tenir un 1 a la variable binària i per la classe k, que es pot estimar de les dades. Es demana:

1. Argumenteu per què aquesta decissió correspon a assumir que les d variables binàries són estadísticament independents *donada la classe*.
2. Doneu l'expressió per les funcions discriminants $a_k(\boldsymbol{x})$ que en resulten. Són discriminants lineals?
3. Doneu l'expressió per la probabilitat a posteriori $P(C_k|\boldsymbol{x})$.

. . . . . .

Problema 4 LDA: model probabilístic generatiu per variables gaussianes [G]

Sabem que un vector aleatori continu en d variables $X = (X_1, \ldots, X_d)^{\mathrm{T}}$ segueix una distribució normal (o gaussiana), cosa que escrivim $X \sim N(\boldsymbol{\mu}, \Sigma)$, quan la seva densitat de probabilitat és:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

on $\boldsymbol{\mu}$ és el vector de les mitjanes i $\Sigma_{d\times d} = (\sigma_{ij}^2)$ és la matriu de covariances. Volem disenyar un classificador probabilístic *generatiu* per un problema de dues classes ($K = 2$), on les distribucions condicionals (per cada classe k) són gaussianes amb **igual** matriu de covariança, o sigui $X_{|C_k} \sim N(\boldsymbol{\mu}_k, \Sigma)$.

1. Mostreu que les probabilitats a posteriori es poden expressar $P(C_k|\boldsymbol{x}) = g(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0)$, on g és la funció *logística*.
2. Doneu (calculant-los amb tots els passos) els valors per $\boldsymbol{w}$ i w_0. Argumenteu si obteniu un classificador lineal o no i per què.
3. Exteneu el resultat al cas d'un número arbitrari de classes $K \geq 2$.

. . . . . .

Problema 5 La fàbrica de píndoles I

La companyia farmacèutica *Nice Pills* ha construit una cinta transportadora que porta dues *classes* de píndoles (adequades per dos tipus de malalties diferents), que anomenem C_1 i C_2. Aquestes píndoles surten en dos colors: $\{yellow, white\}$, que són detectats per una càmera. La companyia fabrica píndoles en proporcions $P(C_1) = \frac{1}{3}, P(C_2) = \frac{2}{3}$. Se'ns facilita també informació sobre la distribució del color per cada classe: $P(yellow|C_1) = \frac{1}{5}, P(white|C_1) = \frac{4}{5}$ i $P(yellow|C_2) = \frac{2}{3}, P(white|C_2) = \frac{1}{3}$. Es demana:

1. Quina és la probabilitat d'error si no s'utilitza el color per classificar?
2. Calcular les probabilitats $P(yellow)$ i $P(white)$ i les probabilitats $P(C_1|yellow)$, $P(C_2|yellow)$, $P(C_1|white)$ i $P(C_2|white)$.

3. Quina és la decissió òptima per pastilles *yellow*? I per pastilles *white*? Quins són els *odds* en ambdós casos?

4. Quina és la probabilitat d'error si s'utilitza el color per classificar? Per què és millor que la de l'apartat 1?

Feu tots els càlculs i doneu tots els resultats en forma de **fraccions**.

. . . . . .

Problema 6 La fàbrica de píndoles II

La companyia farmacèutica *Good Pills* (competidora de l'anterior) ha construit una cinta transportadora que porta dues *classes* de píndoles (adequades per dos tipus de malalties diferents), que anomenem C_1 i C_2. Aquestes píndoles surten en tres colors: $\{yellow, white, red\}$, que són detectats per una càmera. La companyia fabrica píndoles en proporcions $P(C_1) = \frac{1}{3}, P(C_2) = \frac{2}{3}$. Se'ns facilita també informació sobre la distribució del color per cada classe:

	yellow	white	red
C_1	1/5	3/5	1/5
C_2	2/4	1/4	1/4

1. Quina és la probabilitat d'error si no s'utilitza el color per classificar?

2. Calcular les probabilitats $P(yellow)$, $P(white)$ i $P(red)$ i les probabilitats $P(C_1|yellow)$,$P(C_2|yellow)$, $P(C_1|white)$,$P(C_2|white)$,$P(C_1|red)$ i $P(C_2|red)$.

3. Quina és la decissió òptima per pastilles *yellow*? I per pastilles *white*? I per pastilles *red*? Quins són els *odds* en tots els casos?

4. Quina és la probabilitat d'error si s'utilitza el color per classificar? Per què és millor que la de l'apartat 1?

Feu tots els càlculs i doneu tots els resultats en forma de **fraccions**.

. . . . . .

Problema 7 La fàbrica de píndoles III [G]

La companyia farmacèutica *Smart Pills* (competidora de les anteriors) ha construit una cinta transportadora que porta dues *classes* de píndoles (adequades per dos tipus de malalties diferents), que anomenem C_1 i C_2. Aquestes píndoles surten en un ombrejat de colors que va del *yellow* al *white* (que és detectat per una càmera, donant un valor continu en $[0, 2]$). La companyia fabrica píndoles en proporcions $P(C_1) = \frac{1}{3}, P(C_2) = \frac{2}{3}$. Se'ns facilita també informació sobre la distribució (contínua) del color per cada classe:

$$p(x|C_1) = \frac{2-x}{2},\ p(x|C_2) = \frac{x}{2}$$

1. Quina és la probabilitat d'error si no s'utilitza el color per classificar?

2. Calcular la distribució *incondicional* del color $p(x) = P(C_1)p(x|C_1) + P(C_2)p(x|C_2)$.

3. Calcular les distribucions de probabilitat $P(C_1|x)$ i $P(C_2|x)$.

4. Quina és la classificació òptima en funció del color?

5. Quina és la probabilitat d'error si s'utilitza el color per classificar? Per què és millor que la de l'apartat 1?

. . . . . .

Problema 8 Els classificadors LDA i QDA [G]

Sabem que un vector aleatori continu en d variables $X = (X_1, \ldots, X_d)^{\mathrm{T}}$ segueix una distribució normal (o gaussiana), cosa que escrivim $X \sim N(\boldsymbol{\mu}, \Sigma)$, quan la seva densitat de probabilitat és:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

on $\boldsymbol{\mu}$ és el vector de les mitjanes i $\Sigma_{d\times d} = (\sigma_{ij}^2)$ és la matriu de covariances. Volem disenyar un classificador probabilístic *generatiu* per un problema de dues classes ($K = 2$), on les distribucions condicionals (per cada classe k) són gaussianes, és a dir, $X_{|C_k} \sim N(\boldsymbol{\mu}_k, \Sigma_k)$. Sabem que el classificador de mínim risc (anomenat *regla de Bayes*) que minimitza la probabilitat d'error s'obté amb la fòrmula de Bayes:

$$P(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)P(C_k)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|C_k)P(C_k)}{p(\boldsymbol{x}|C_1)P(C_1) + p(\boldsymbol{x}|C_2)P(C_2)}$$

triant-se la classe $C_k, k = 1, 2$ que maximitza aquestes probabilitats a posteriori. Es demana:

1. Construïu la funció discriminant per la classe C_k com $g_k(\boldsymbol{x}) = \ln\{P(C_k)p(\boldsymbol{x}|C_k)\}$; elimineu termes que no afectin el resultat. Argumenteu quin tipus de superfícies de separació en resulten.

2. Assumim ara que totes les classes ténen **igual** matriu de covariança, o sigui $X_{|C_k} \sim N(\boldsymbol{\mu}_k, \Sigma)$. Simplifiqueu l'expressió anterior al màxim. Argumenteu quin tipus de superfícies de separació en resulten. Pel cas $K = 2$, és usual construir un únic discriminant $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$ (sovint anomenat un *dichotomizer*). Expresseu-lo.

3. Veiem-ne un petit exemple numèric per $d = 3$. Suposem les densitats gaussianes de classe:

$$\boldsymbol{\mu}_1 = (0,0,0)^{\mathrm{T}}, \boldsymbol{\mu}_2 = (1,1,1)^{\mathrm{T}}, \Sigma_1 = \Sigma_2 = diag\Big(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}\Big), P(C_2) = 2P(C_1)$$

   Construïu el *dichotomizer* i apliqueu-lo a la predicció de l'exemple de *test* $\boldsymbol{x}^* = (0.1, 0.7, 0.8)^{\mathrm{T}}$

. . . . . .

Problema 9

Considerem dues distribucions condicionals (per cada classe) són gaussianes bivariades ($d = 2$) amb **igual** matriu de covariança, de la forma $X_{|C_k} \sim N(\boldsymbol{\mu}_k, \Sigma)$, $k = 1, 2$.

1. Suposant que les dues classes són igual de probables, calculeu la regla de classificació òptima.

2. Apliqueu el resultat a les dades $\mu_1 = (0,0)^{\mathrm{T}}$, $\mu_2 = (3,3)^{\mathrm{T}}$ i $\Sigma = \left(\begin{smallmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{smallmatrix}\right)$, per obtenir una regla de classificació concreta.

3. Classifiqueu el punt $\boldsymbol{x}^* = (1.0, 2.2)^{\mathrm{T}}$.

4. Calculeu $\|\boldsymbol{x}^* - \boldsymbol{\mu}_1\|$ i $\|\boldsymbol{x}^* - \boldsymbol{\mu}_2\|$ i notareu que $\|\boldsymbol{x}^* - \boldsymbol{\mu}_1\| > \|\boldsymbol{x}^* - \boldsymbol{\mu}_2\|$. Com quadra això amb el resultat del punt anterior?

. . . . . .

Problema 10 Juguem a tennis?

Dos amics han recopilat dades sobre diverses vegades en que havien quedat per jugar a tennis (unes vegades van acabar jugant i altres no, depenent de les previsions meteorològiques).

Outlook	Temperature	Humidity	Wind	PlayTennis?
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Construïu un classificador Naïve Bayes i utilitzeu-lo per determinar si haurien de jugar a tenis en les condicions d'un exemple de *test* $\boldsymbol{x}^* = (\text{Sunny, Hot, Normal, Weak})^{\mathrm{T}}$. Noteu que no cal calcular totes les probabilitats possibles, sinó només les imprescindibles per aquesta predicció concreta.

. . . . . .

Problema 11 Interpretació de models de regressió logística

Considerem un model de regressió logística $y(\boldsymbol{x}) = g(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0)$, on g és la funció *logística*. Es demana:

1. Deriveu una interpretació per un coeficient qualsevol w_i (diferent de w_0) a partir de la variació dels *odds* quan x_i passa a ser $x_i + \delta_i$ i apliqueu-la al cas particular $\delta_i = 1$.

2. Tenim $y(\boldsymbol{x}) = g(1.3x_1 + 0.7x_2 - 0.29x_3 + 0.54)$. Apliqueu la interpretació al coeficient de x_1 quan $\delta_1 = 1$ i al coeficient de x_3 quan $\delta_3 = -0.5$.

. . . . . .

Problema 12 Obtenció de la regressió logística

Una manera elegant d'arribar al model de regressió logística és partir dels *odds*. En tasques de classificació binària (dues classes, C_1 i C_2), considerem el logaritme natural dels *odds* (anomenat *logit* or *log-odds*) per un $\boldsymbol{x}$ qualsevol:

$$\ln\left(\frac{P(C_1|\boldsymbol{x})}{P(C_2|\boldsymbol{x})}\right) = \ln\left(\frac{P(C_1|\boldsymbol{x})}{1 - P(C_1|\boldsymbol{x})}\right)$$

Resoleu aquesta fòrmula en la probabilitat, calculant la funció inversa de la *logit*. Definiu el model que en resulta com el de regressió logística i doneu-ne una interpretació en termes de linealitat del model.

. . . . . .