Classification Bayesienne et Bayésien Naïf

Alexandre Allauzen
allauzen@limsi.fr

Université Paris Sud — LIMSI

Septembre 2017

Plan

Le cadre

- Soit une VA X une observation à classer
- Soit une VA Y désignant la classe à affecter à X

Classification

La VA Y est discrète par définition.

Objectif : trouver la règle de décision

- permettant d'affecter y à x
- x et y sont des réalisations de X et Y
- Utilisons les distributions de probabilités sur X et Y, estimées sur les données d'apprentissage

Décision à priori

Classification binaire

$\mathcal{A}_Y = \{0, 1\}$

$$\text{Décider } y = 1 \text{ si } P(Y = 1) > P(Y = 0)$$

Multi-classe

$\mathcal{A}_Y = \{1, ..., K\}$

$$\text{Décider } y = \text{argmax}_k\, P(Y = k)$$

Observons avant de décider

L'objet à classer est représenté par un ensemble de caractéristiques :

$$\boldsymbol{x} = (x_1, ..., x_d)$$

réalisation du vecteurs de V.As :

$$\boldsymbol{X} = (X_1, ...X_d)$$

Chaque caractéristique est une V.A et

$$P(\boldsymbol{X} = x) = P(X_1 = x_1, ..., X_d = x_d)$$

Notion de risque (cas général)

Notations

- Les caractéristiques sont regroupées dans le vecteur de VA $\boldsymbol{X} = \boldsymbol{x}$
- La classe $y \in \mathcal{A}_Y = \{1, ..., K\}$
- La décision α_i : affecté $\boldsymbol{x}$ à la classe $y = i$
- Une fonction de perte $\lambda(\alpha_i|j)$ de décider α_i alors que la décision juste est j

Expected (conditional) Risk

Observons $\boldsymbol{x}$, quel est l'espérance du risque de décider α_i :

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{K} \lambda(\alpha_i|j)P(Y = j|\boldsymbol{x})$$

- Pour chaque classe j, c'est la bonne réponse avec comme probabilité $P(Y = j|\boldsymbol{x})$
- Et le coût de décider α_i est $\lambda(\alpha_i|j)$

Minimisons le risque

Zero-one loss

$$\lambda(\alpha_i|j) = 0 \text{ si } i = j$$
$$\lambda(\alpha_i|j) = 1 \text{ si } i \neq j$$

Expected Risk

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{K} \lambda(\alpha_i|j) P(Y = j|\boldsymbol{x}) \qquad = \sum_{j=1, j \neq i}^{K} \lambda(\alpha_i|j) P(Y = j|\boldsymbol{x}) = 1 - P(Y = i|\boldsymbol{x})$$

Donc choisir la décision de risque minimum revient à choisir $y = j$ tel que $P(Y = j|\boldsymbol{x})$ soit maximum !

Décision MAP (Maximum a posteriori)

Classification binaire : $\mathcal{A}_Y = \{0, 1\}$

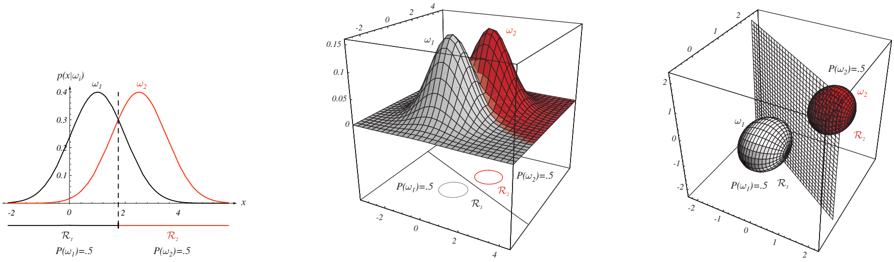
$$\text{Décider } y = 1 \text{ si } P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) > P(Y = 0|\boldsymbol{X} = \boldsymbol{x})$$
$$\text{si } \frac{P(\boldsymbol{X} = \boldsymbol{x}|Y = 1)}{P(\boldsymbol{X} = \boldsymbol{x}|Y = 0)} > \frac{P(Y = 0)}{P(Y = 1)}$$

Pour des classes équilibrées : $y = 1$ si $P(\boldsymbol{X} = \boldsymbol{x}|Y = 1) > P(\boldsymbol{X} = \boldsymbol{x}|Y = 0)$

Multi-classe : $\mathcal{A}_Y = \{1, ..., K\}$

$$\text{Décider } y = \operatorname*{argmax}_k P(Y = k|\boldsymbol{X} = \boldsymbol{x})$$

Surface de décision



From Duda and Hart

D'autres pertes

Les faux billets

Le problème des faux billets

2 classes

- $y = 1$: vrai billet, $P(Y = 1) = 0.6$
- $y = 0$: faux billet, $P(Y = 0) = 0.4$

2 actions

- α_1 accepter un billet de 100
- α_0 refuser un billet de 100

Fonction de coût

- $\lambda(\alpha_1|1) = 1$ accepter un vrai billet
- $\lambda(\alpha_1|0) = 101$ accepter un faux billet
- $\lambda(\alpha_0|1) = 11$ refuser un vrai billet
- $\lambda(\alpha_0|0) = 1$ refuser un faux billet

Les faux billets - 2

Le risque conditionnel

Posons $\lambda_{ij} = \lambda(\alpha_i|j)$, et

$$
\begin{aligned}
R(\alpha_1|\boldsymbol{x}) &= \lambda_{11}P(Y=1|\boldsymbol{x}) + \lambda_{10}P(Y=0|\boldsymbol{x}) \\
R(\alpha_0|\boldsymbol{x}) &= \lambda_{01}P(Y=1|\boldsymbol{x}) + \lambda_{00}P(Y=0|\boldsymbol{x})
\end{aligned}
$$

Écrire la règle de décision :

Décider $Y = 1$ si

$$
R(\alpha_1|\boldsymbol{x}) < R(\alpha_0|\boldsymbol{x})
$$
$$
(\lambda_{11} - \lambda_{01})P(Y=1|\boldsymbol{x}) < (\lambda_{00} - \lambda_{10})P(Y=0|\boldsymbol{x})
$$
$$
(\lambda_{01} - \lambda_{11})P(Y=1|\boldsymbol{x}) > (\lambda_{10} - \lambda_{00})P(Y=0|\boldsymbol{x})
$$
$$
\frac{p(\boldsymbol{x}|Y=1)}{p(\boldsymbol{x}|Y=0)} > \frac{\lambda_{00} - \lambda_{10}}{\lambda_{11} - \lambda_{01}} \frac{P(Y=0)}{P(Y=1)}
$$

Décision MAP - 2

La décision optimale

- On minimise le risque d'erreur
- mais dans le cas idéal (on connait les distributions).
- En pratique, ces distributions sont estimées.

Ce qu'il reste à faire

Estimer les distribution !

- Estimer directement $P(Y|\boldsymbol{X})$: modèle discriminant
- Appliquer la formule de Bayes : modèle génératif

$$
\begin{aligned}
y &= \operatorname*{argmax}_k P(Y = k|\boldsymbol{X} = \boldsymbol{x}) & &= \operatorname*{argmax}_k \frac{P(\boldsymbol{X} = \boldsymbol{x}|Y = k)P(Y)}{P(\boldsymbol{X} = \boldsymbol{x})} \\
&= \operatorname*{argmax}_k P(\boldsymbol{X} = \boldsymbol{x}|Y = k)P(Y) & &= \operatorname*{argmax}_k P(\boldsymbol{X} = \boldsymbol{x}, Y = k)
\end{aligned}
$$

Plan

Un modèle génératif

Modélisation de $P(\boldsymbol{X}, Y) = P(\boldsymbol{X}|Y)P(Y)$:

- $P(Y)$: facile !
- Reste à paramétriser et estimer $P(\boldsymbol{X}|Y)$

Complexité

- 100 caractéristiques gaussienne → 10100 paramètres par classe
- 784 caractéristiques gaussienne → 615440 paramètres par classe
- 100 caractéristiques binaires → $1.26e30$ paramètres par classe

Hypthèse Naïve

Les composantes de $\boldsymbol{X}$ sont indépendantes les unes des autres :

$$P(\boldsymbol{X}|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

Gaussian Bayes Naive

- Chaque composante x_i de $\boldsymbol{x}$ est réelle
- $X_i|Y = y \sim \mathcal{N}(\mu_{i,y}, \sigma_{i,y})$ (matrice de covariance diagonale, un vecteur)
- Nombre de paramètres : $2Kd + K$
- Estimation : regrouper les $\boldsymbol{x}$ par classe, puis calculer moyenne et variance par composante
- Inférence :

$$P(Y = y|\boldsymbol{X} = \boldsymbol{x}) \propto P(Y = y) \prod_{i=1}^{d} \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma_{i,y}^2}} \times e^{-\frac{1}{2}\frac{(x_i - \mu_{i,y})^2}{\sigma_{i,y}^2}}$$

$$\log P(Y = y|\boldsymbol{X} = \boldsymbol{x}) \propto \log P(Y = y) - \frac{1}{2} \sum_{i=1}^{d} \left(\log(2 \cdot \pi \cdot \sigma_{i,y}^2) + \frac{(x_i - \mu_{i,y})^2}{\sigma_{i,y}^2} \right)$$

Exemple : classification de texte

Applications

- Filtrage de mails : $Y = \{spam, ham\}$
- Filtrage de mails : $Y = \{pro_{urgent}, pro_{plus\ tard}, perso_{urgent}, pro_{plus\ tard}, spam\}$
- Sentiments : $Y = \{negatif, positif\}$ ou $Y = \{negatif, neutre, positif\}$

Choix de représentation : $\boldsymbol{x}$

this movie is just great , with a great music , while a bit long

$\mathcal{V}$(vocabulaire)	sac binaire	compte
the	0	0
awesome	0	0
this	1	1
long	1	1
great	1	2
...	...	...
modèle	Bernoulli	multinomial

Modèle de Bernoulli

à pile où face

Pour chaque mot de $\mathcal{V}$ indicé par i :

- X_i désigne la présence du mot i
- $\mathcal{A}_{X_i} = \{0, 1\}$ et $P(X_i = 1) = \pi_i$

$\mathcal{V}$(vocabulaire)	$\boldsymbol{x}$	$P(\boldsymbol{X} = \boldsymbol{x})$
the	0	$P(X_1 = 0) = 1 - \pi_1$
awesome	0	$P(X_2 = 0) = 1 - \pi_2$
this	1	$P(X_3 = 1) = \pi_3$
long	1	$P(X_4 = 1) = \pi_4$
great	1	$P(X_5 = 1) = \pi_5$

$$P(\boldsymbol{X} = [0, 0, 1, 1, 1]) = (1 - \pi_1)(1 - \pi_2)\pi_3\pi_4\pi_5,$$

Remarque : ne pas confondre avec le modèle binomial !

Binomial Bayes Naive

Inférence : pour chaque classe $y \in \mathcal{A}_Y$:

$$P(Y=y|\boldsymbol{X}=\boldsymbol{x}) \propto P(\boldsymbol{X}=\boldsymbol{x}|Y=y)P(Y=y) = \Big(\prod_{i=1}^{d} \pi_{i,y}^{x_i}(1-\pi_{i,y})^{(1-x_i)}\Big)P(Y=y)$$

- Nombre de paramètres : $K \times d + K$
- Estimation :

$$\pi_{i,y} = \frac{\text{nombre de doc dans la classe y contenant le mot i}}{\text{nombre de doc de la classe y}} = \frac{n(i,y)}{n(y)}$$

Inférence en log :

$$\begin{aligned}\log(P(Y=y|\boldsymbol{X}=\boldsymbol{x})) &= \log(P(\boldsymbol{X}=\boldsymbol{x}|Y=y)) + \log(P(Y=y)) \\ &= \sum_{i=1}^{d} x_i \log(\pi_{i,y}) + (1-x_i)\log(1-\pi_{i,y}) + \log(P(Y=y))\end{aligned}$$

Lissage

$$\text{Si } \frac{n(i,y)}{n(y)} = 0 \Rightarrow P(X_i = 1|Y) = 0$$

$$\text{Si } \frac{n(i,y)}{n(y)} = 1 \Rightarrow P(X_i = 0|Y) = 0$$

Dans les deux cas on peut avoir $P(Y = y|\boldsymbol{X} = \boldsymbol{x}) = 0$

Lissage de Laplace

Soit $\alpha > 0$

$$\pi_{i,y} = \frac{n(i,y) + \alpha}{n(y) + 2\alpha}$$

Modèle multinomial

modéliser les comptes

- Soit la VA X : l'apparition d'un mot parmi $\mathcal{V}$dans le texte
- $\mathcal{A}_X = \mathcal{V} = \{the, awesome, this, long, great, ...\}$
- Les paramètres : $\{\beta_{the}, \beta_{awesome}, \beta_{this}, \beta_{long}, \beta_{great}, ...\}$
- $P(X = the) = \beta_{the}$

Histoire générative :

- Choisir le nombre d'occurences de mots L du document
- Pour chaque occurence : tirer un mot selon la distribution catégorielle de X

La probabilité d'un document représenté par $\boldsymbol{x}$ réalisation de $\boldsymbol{X}$:

$$P(\boldsymbol{X} = \boldsymbol{x}) = K(L, d) \prod_{i=1}^{d} \beta_i^{x_i} \propto \prod_{i=1}^{d} \beta_i^{x_i}$$

$$\log(P(\boldsymbol{X} = \boldsymbol{x})) = \log(K(L, d)) + \sum_{i=1}^{d} x_i \log \beta_i,$$

avec $K(L, d)$ la fonction de normalisation, considérée comme une constante dans de nombreux contextes.

Modèle multinomial - 2

$\mathcal{V}$(vocabulaire)	$\boldsymbol{x}$: comptes	$P(\boldsymbol{X} = \boldsymbol{x})$
the	0	$\beta^0_{the} = 1$
awesome	0	$\beta^0_{awesome} = 1$
this	1	β^1_{this}
long	1	β^1_{long}
great	2	β^2_{great}
...	...	...

- Les mots de $\mathcal{V}$, absents du document, n'intervennent pas.
- Le nombre de paramètres pour une disrtibution : d
- $\sum_{i=1}^{d} \beta_i = 1$
- Modèle de document de la même taille

Multinomial Bayes Naive

Inférence : pour chaque classe $y \in \mathcal{A}_Y$:

$$P(Y=y|\boldsymbol{X}=\boldsymbol{x}) \propto P(\boldsymbol{X}=\boldsymbol{x}|Y=y)P(Y=y) = \Big(\prod_{i=1}^{d} \beta_{i,y}^{x_i}\Big)P(Y=y)$$

- Nombre de paramètres : $K \times d + K$
- Estimation :

$$\beta_{i,y} = \frac{\text{nombre d'occurence du mot i dans les docs de la classe y}}{\text{nombre total de mots dans les docs de la classe y}} = \frac{c(i,y)}{c(y)}$$

Inférence en log :

$$\begin{aligned}\log(P(Y=y|\boldsymbol{X}=\boldsymbol{x})) &= \log(P(\boldsymbol{X}=\boldsymbol{x}|Y=y)) + \log(P(Y=y)) \\ &= \sum_{i=1}^{d} x_i \log(\beta_{i,y}) + \log(P(Y=y))\end{aligned}$$

Lissage

$$\text{Si } \frac{n(i,y)}{n(y)} = 0 \Rightarrow P(X_i = 1|Y) = 0$$

Alors on peut avoir $P(Y = y|\boldsymbol{X} = \boldsymbol{x}) = 0$

Lissage de Laplace

Soit $\alpha > 0$

$$\beta_{i,y} = \frac{c(i,y) + \alpha}{c(y) + d\alpha}$$

Résumé : classification Bayésienne

Décision (inférence)

$\mathcal{A}_Y = \{1, ..., K\}$

$$\text{Décider } y = \text{argmax}_k \, P(Y = k|\boldsymbol{X} = \boldsymbol{x})$$

Estimation des paramètres (apprentissage)

- Choix de la paramètrisation de $P(Y = k|\boldsymbol{X} = \boldsymbol{x})$

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{X} = \boldsymbol{x}, Y = Y)}{P(\boldsymbol{X} = \boldsymbol{x})} \propto P(\boldsymbol{X} = \boldsymbol{x}|Y = Y)P(Y = y)$$

- Choix de la paramètrisation de $P(\boldsymbol{X} = \boldsymbol{x}|Y = k)$

Résumé : Bayésien Naïf

Hypothèse simplificatrice : les composantes de $\boldsymbol{X}$ sont indépendantes connaissant Y

$$P(\boldsymbol{X} = \boldsymbol{x}|Y = Y) = \prod_{i=1}^{d} P(X_i = x_i|Y = Y)$$

- En continu : $P(X_i = x_i|Y = Y) \sim \mathcal{N}(\mu_i, \sigma_i)$
- En discret : $P(X_i = x_i|Y = Y)$ est binomiale ou multinomiale