Introduction au Machine Learning

C1 - Premiers modèles pour la classification

Lucas Gnecco Heredia

LAMSADE - Université Paris-Dauphine
lucas.gnecco-heredia@dauphine.psl.eu - C602

13 septembre 2022

* *Remerciement spécial à Florian Yger*

Plan

1. Introduction / Rappel
2. k-plus proches voisins (k-ppv) (IML Ch 8), (ESL Sec 13.3)
3. Classification naïve bayésienne (IML Ch 4, Sec 4.4)

Motivation

Exemples pratiques

classification de chiffres manuscrits (MNIST, USPS,...)
output : classes $0, 1, \cdots, 9$ - y
input : images x

Question à résoudre

Motivation

Exemples pratiques

classification de chiffres manuscrits (MNIST, USPS,...)
output : classes $0, 1, \cdots, 9$ - y
input : images x

Question à résoudre

- Comment reconnaître un chiffre manuscrit à partir de son image ?

Motivation

Exemples pratiques

classification de chiffres manuscrits (MNIST, USPS,...)
output : classes $0, 1, \cdots, 9$ - y
input : images x

Question à résoudre

- Comment reconnaître un chiffre manuscrit à partir de son image ?
- Comment trouver la classe d'une image ?

Approches possibles

Approche naïve

Enumération d'un ensemble de règles

- ex : si l'intensité du pixel à la position (i, j) est plus grand que le seuil τ et si ...
  alors c'est un "3"

Approches possibles

Approche naïve

Enumération d'un ensemble de règles

- ex : si l'intensité du pixel à la position (i, j) est plus grand que le seuil τ et si ...
  alors c'est un "3"
- trop fastidieux

Approches possibles

Approche naïve

Enumération d'un ensemble de règles

- ex : si l'intensité du pixel à la position (i, j) est plus grand que le seuil τ et si ...
  alors c'est un "3"
- trop fastidieux
- difficile de couvrir tous les cas possibles

Approches possibles

Approche naïve

Enumération d'un ensemble de règles

- ex : si l'intensité du pixel à la position (i, j) est plus grand que le seuil τ et si ...
  alors c'est un "3"
- trop fastidieux
- difficile de couvrir tous les cas possibles
- pas assez robuste aux variations dans les données

Approches possibles

Approche naïve

Enumération d'un ensemble de règles

- ex : si l'intensité du pixel à la position (i, j) est plus grand que le seuil τ et si ...
  alors c'est un "3"
- trop fastidieux
- difficile de couvrir tous les cas possibles
- pas assez robuste aux variations dans les données

Approche moderne

Donner la capactité à l'ordinateur d'apprendre à partir des données

Approches possibles

Approche naïve

Enumération d'un ensemble de règles

- ex : si l'intensité du pixel à la position (i, j) est plus grand que le seuil τ et si ...
  alors c'est un "3"
- trop fastidieux
- difficile de couvrir tous les cas possibles
- pas assez robuste aux variations dans les données

Approche moderne

Donner la capactité à l'ordinateur d'apprendre à partir des données
Type d'approches développées en ML (et que nous allons étudier)

Apprentissage artificiel

Principe général

Un algorithme d'apprentissage procède comme suit :

- parcours des données d'entraînement (*training data*)
- création d'un "programme" capable de généraliser à de nouvelles données

Apprentissage artificiel

Principe général

Un algorithme d'apprentissage procède comme suit :

- parcours des données d'entraînement (*training data*)
- création d'un "programme" capable de généraliser à de nouvelles données

Formellement, ce programme est l'implémentation d'un modèle qui à chaque point prédit sa classe/label (i.e. $\hat{y}_i = f(x_i)$)

Apprentissage artificiel

Principe général

Un algorithme d'apprentissage procède comme suit :

- parcours des données d'entraînement (*training data*)
- création d'un "programme" capable de généraliser à de nouvelles données

Formellement, ce programme est l'implémentation d'un modèle qui à chaque point prédit sa classe/label (i.e. $\hat{y}_i = f(x_i)$)

Remarque

On retrouve un cadre connnu mais à la différence de la régression, $f(x)$ et y prennent leurs valeurs dans un ensemble discret (et sans relation d'ordre) - l'ensemble des classes-.

Cadre pratique

Ensemble d'apprentissage/entraînement

- constitué de données pour lesquelles le label est connu.
- représentatif des données futures (c-à-d, de l'ensemble de test)

On considère que les ensembles d'apprentissage et de test sont constitués d'individus tirés au hasard dans la population à modéliser [1] (même distributions de probabilité).

1. Il existe des approches ne faisant pas cette hypothèse mais on est alors dans un cas très particulier.

Modélisation

- f doit permettre de d'approximer une relation d'entrée-sortie (entre x et y) à partir des données
- cette fonction de décision doit pouvoir faire de bonnes prédictions sur des données inconnues.
  On appelle cette capacité, la généralisation, par opposition au sur-apprentissage.

Analogie

un élève apprenant une règle de grammaire à partir d'exemples

- l'apprentissage "par coeur" est inefficace car ne permet pas de s'adapter à de nouveaux cas (sur-apprentissage)
- l'élève déduit une règle générale dans le but de pouvoir l'appliquer plus tard (généralisation).
- il y a un compromis entre la simplicité de la règle et sa généralisation (on parle aussi de complexité pour une fonction de décision).

1 Introduction / Rappel

2 k-plus proches voisins (k-ppv) (IML Ch 8), (ESL Sec 13.3)

3 Classification naïve bayésienne (IML Ch 4, Sec 4.4)

Principe

En anglais

K-Nearest neighbours (K-NN)

Principe

En anglais

K-Nearest neighbours (K-NN)

Caractéristiques

- techniquement très simple
- très peu d'apprentissage en tant que tel (*lazy learner*)
  s'apparente à un *raisonnement par cas* - agissant en fonction des choix effectués dans des cas similaires

Principe

Fonctionnement

- le classement de chaque individu inconnu s'opère en regardant la classe des k individus les plus proches (ses voisins) dans l'ensemble d'apprentissage en en choisissant la classe la plus représentée. On combine les *opinions* de plusieurs voisins.
- l'hyperparamètre k est choisi pour faire le meilleur classement possible (sur l'ensemble des données à prédire)

Un peu plus formellement

En notant $\mathcal{N}_k(x)$ l'ensemble des k plus proches voisins de x (dans l'ensemble d'apprentissage $\mathcal{D}$:

$$f(x) = \text{argmax}_c \sum_{i/x_i \in \mathcal{N}_k(x)} \delta(y_i, c)$$

Ingrédients essentiels

On a besoin de :

- un ensemble d'observations labelisées

Ingrédients essentiels

On a besoin de :

- un ensemble d'observations labelisées
- une distance (ou une mesure de similarité) entre observations [2]

2. permet de traiter tout type de données - graphes, images, etc...- dès qu'une distance peut être définie.

Ingrédients essentiels

On a besoin de :

- un ensemble d'observations labelisées
- une distance (ou une mesure de similarité) entre observations [2]
- une valeur de k

2. permet de traiter tout type de données - graphes, images, etc...- dès qu'une distance peut être définie.

Classement d'une nouvelle observation

Algo

- Calcul de la distance entre cette observation et toutes les observations connues
- identification des k plus proches voisin
- utilisation de la clases des voisins pour déterminer celle de l'observation inconnue par un votre majoritaire (éventuellement pondéré par l'inverse de la distance)

Complexité

Quelle complexité pour chaque étape ?

Classement d'une nouvelle observation

Algo

- Calcul de la distance entre cette observation et toutes les observations connues
- identification des k plus proches voisin
- utilisation de la clases des voisins pour déterminer celle de l'observation inconnue par un votre majoritaire (éventuellement pondéré par l'inverse de la distance)

Complexité

Quelle complexité pour chaque étape ?

Classement d'une nouvelle observation

Algo

- Calcul de la distance entre cette observation et toutes les observations connues $\mathcal{O}(n)$
- identification des k plus proches voisin $\mathcal{O}(n \log(n))$
- utilisation de la clases des voisins pour déterminer celle de l'observation inconnue par un votre majoritaire (éventuellement pondéré par l'inverse de la distance) $\mathcal{O}(k)$

Complexité

Quelle complexité pour chaque étape ?

Hyperparamètre k

Impact du choix de la valeur

Hyperparamètre k

Impact du choix de la valeur

- k trop petit : forte sensibilité aux valeurs abberrantes

Hyperparamètre k

Impact du choix de la valeur

- k trop petit : forte sensibilité aux valeurs abberrantes
- k trop grand : le voisinage peut inclure des points d'autres classes

Hyperparamètre k

Impact du choix de la valeur

- k trop petit : forte sensibilité aux valeurs abberrantes
- k trop grand : le voisinage peut inclure des points d'autres classes
- plus k grandit, plus la frontière de décision devient lisse et régulière.

Hyperparamètre k

Impact du choix de la valeur

- k trop petit : forte sensibilité aux valeurs abberrantes
- k trop grand : le voisinage peut inclure des points d'autres classes
- plus k grandit, plus la frontière de décision devient lisse et régulière.
- quand $k = n$, la classe majoritaire est tout le temps prédite

En pratique

- choix de k par validation croisée (cf chapitre de méthodologie)
- heuristique quelquefois choisie : $k = \sqrt{n}$ ou $k = \sqrt{n/C}$ (i.e. le nombre moyen de points par classes)

Un autre hyperparamètre à considérer : la distance

Dans le cadre d'un 1-ppv (par soucis de simplicité)

Quelques exemples de distances (pour des données non-structurées)

Un autre hyperparamètre à considérer : la distance

Dans le cadre d'un 1-ppv (par soucis de simplicité)

Quelques exemples de distances (pour des données non-structurées)

- $d_2(x, y) = ||x-y||_2 = \sqrt{(x-y)^\top (x-y)} = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$ (euclidienne)

Un autre hyperparamètre à considérer : la distance

Dans le cadre d'un 1-ppv (par soucis de simplicité)

Quelques exemples de distances (pour des données non-structurées)

- $d_2(x, y) = ||x-y||_2 = \sqrt{(x-y)^\top (x-y)} = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$ (euclidienne)
- $d_1(x, y) = ||x-y||_1 = \sum_{i=1}^m |x_i - y_i|$ (Manhattan)

Un autre hyperparamètre à considérer : la distance

Dans le cadre d'un 1-ppv (par soucis de simplicité)

Quelques exemples de distances (pour des données non-structurées)

- $d_2(x, y) = ||x-y||_2 = \sqrt{(x-y)^\top (x-y)} = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$ (euclidienne)
- $d_1(x, y) = ||x-y||_1 = \sum_{i=1}^m |x_i - y_i|$ (Manhattan)
- $d_\infty(x, y) = ||x-y||_\infty = \max_i |x_i - y_i|$ (Minkowski)

Un autre hyperparamètre à considérer : la distance

Dans le cadre d'un 1-ppv (par soucis de simplicité)

Quelques exemples de distances (pour des données non-structurées)

- $d_2(x, y) = ||x-y||_2 = \sqrt{(x-y)^\top (x-y)} = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$ (euclidienne)
- $d_1(x, y) = ||x-y||_1 = \sum_{i=1}^m |x_i - y_i|$ (Manhattan)
- $d_\infty(x, y) = ||x-y||_\infty = \max_i |x_i - y_i|$ (Minkowski)
- $d_p(x, y) = ||x-y||_p = \sqrt[p]{\sum_{i=1}^m (x_i - y_i)^p}$

Un autre hyperparamètre à considérer : la distance

Dans le cadre d'un 1-ppv (par soucis de simplicité)

Quelques exemples de distances (pour des données non-structurées)

- $d_2(x,y) = ||x-y||_2 = \sqrt{(x-y)^\top (x-y)} = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$ (euclidienne)
- $d_1(x,y) = ||x-y||_1 = \sum_{i=1}^m |x_i - y_i|$ (Manhattan)
- $d_\infty(x,y) = ||x-y||_\infty = \max_i |x_i - y_i|$ (Minkowski)
- $d_p(x,y) = ||x-y||_p = \sqrt[p]{\sum_{i=1}^m (x_i - y_i)^p}$
- $d_\Sigma(x,y) = ||x-y||_\Sigma = \sqrt{(x-y)^\top \Sigma (x-y)}$ - avec $\Sigma \succcurlyeq 0$ - (Mahalanobis)

Un autre hyperparamètre à considérer : la distance

Dans le cadre d'un 1-ppv (par soucis de simplicité)

Quelques exemples de distances (pour des données non-structurées)

- $d_2(x,y) = ||x-y||_2 = \sqrt{(x-y)^\top (x-y)} = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$ (euclidienne)
- $d_1(x,y) = ||x-y||_1 = \sum_{i=1}^m |x_i - y_i|$ (Manhattan)
- $d_\infty(x,y) = ||x-y||_\infty = \max_i |x_i - y_i|$ (Minkowski)
- $d_p(x,y) = ||x-y||_p = \sqrt[p]{\sum_{i=1}^m (x_i - y_i)^p}$
- $d_\Sigma(x,y) = ||x-y||_\Sigma = \sqrt{(x-y)^\top \Sigma (x-y)}$ - avec $\Sigma \succcurlyeq 0$ - (Mahalanobis)

Comparaison des boules unitaires

Limites de l'approche

Limites de l'approche

- apprentissage paresseux (*lazy learner*) - pas d'apprentissage

Limites de l'approche

- apprentissage paresseux (*lazy learner*) - pas d'apprentissage
- pas de construction explicite d'un modèle (méthode transductive)

Limites de l'approche

- apprentissage paresseux (*lazy learner*) - pas d'apprentissage
- pas de construction explicite d'un modèle (méthode transductive)
- nécessite de manipuler tout l'ensemble d'apprentissage à chaque prédiction

Limites de l'approche

- apprentissage paresseux (*lazy learner*) - pas d'apprentissage
- pas de construction explicite d'un modèle (méthode transductive)
- nécessite de manipuler tout l'ensemble d'apprentissage à chaque prédiction
- nécessite une grande capacité de stockage et de calcul

Limites de l'approche

- apprentissage paresseux (*lazy learner*) - pas d'apprentissage
- pas de construction explicite d'un modèle (méthode transductive)
- nécessite de manipuler tout l'ensemble d'apprentissage à chaque prédiction
- nécessite une grande capacité de stockage et de calcul
- fonctionne mieux en faible dimension

Limites de l'approche

- apprentissage paresseux (*lazy learner*) - pas d'apprentissage
- pas de construction explicite d'un modèle (méthode transductive)
- nécessite de manipuler tout l'ensemble d'apprentissage à chaque prédiction
- nécessite une grande capacité de stockage et de calcul
- fonctionne mieux en faible dimension
- performances très dépendante du choix de k et de distance

Variantes I

k-ppv pour la régression

$$f(x) = \frac{1}{k} \sum_{i/x_i \in \mathcal{N}_k(x)} y_i$$

On prédit pour x, la moyenne des y_i de ses voisins.

Variantes II

ϵ-voisins

On considère des boules (de diamètre ϵ) à la place d'un voisinage. Limite : en absence d'exemple dans la boule, la décision n'est pas définie.

Pondération des voisins

On donne plus de crédit à un voisin proche et on peut utiliser commme poids :

- $w_i = \frac{1}{1+d(x,x_i)}$
- $w_i = e^{-(\frac{1}{2}d(x,x_i))}$

Variantes II

ϵ-voisins

On considère des boules (de diamètre ϵ) à la place d'un voisinage. Limite : en absence d'exemple dans la boule, la décision n'est pas définie.

Pondération des voisins

On donne plus de crédit à un voisin proche et on peut utiliser commme poids :

- $w_i = \frac{1}{1+d(x,x_i)}$
- $w_i = e^{-(\frac{1}{2}d(x,x_i))}$
- une autre fonctione décroissante (monotone) de la distance

Application (filtrage collaboratif)

Dans un système de recommandation, soit :

- $r(u, a)$ la note donnée par l'utilisateur u à l'objet a.
- $s(a, b)$ une similarité [3] entre deux objets.
- $\mathcal{N}_u^k(a)$ les k plus proches voisins de l'objet a parmi ceux notés par u

On pourra donner comme recommandation à un utilisateur u en cherchant u qui maximise :

$$f(u, a) = \frac{\sum_{b \in \mathcal{N}_u^k(a)} s(a, b) r(u, b)}{\sum_{b \in \mathcal{N}_u^k(a)} |s(a, b)|}$$

3. Définie comme la similarité des notes données par les utilisateurs pour les objets a et b

TD - exercice 1 - les dangers de labelisation en ligne

On dispose de l'ensemble d'apprentissage suivant :

id	x	y
1	(1 ;3)	A
2	(6 ;5)	B
3	(8 ;3)	B
4	(2 ;4)	A

et des données de test :

id	x	y
5	(6 ;3)	?
6	(2 ;2)	?
7	(8 ;1)	?
8	(4 ;3)	?
9	(4 ;4)	?
10	(5 ;2)	?
11	(7 ;4)	?
12	(3 ;1)	?
13	(8 ;5)	?
14	(3 ;2)	?

TD - exercice 1 - les dangers de labelisation en ligne

1. représenter ce jeu de données
2. en utilisant un 3-ppv, prédire itérativement les labels de points 5 à 13 (données de test) en ajoutant les prédictions à l'ensemble d'apprentissage (au fur et à mesure)
3. montrer que le résultat dépend de l'ordre de présentation des exemples.

TD (extrait de "introduction au ML" de C. Azencott)

1. On a le jeu de données de classification binaire suivant :

$x^{(1)}$	1	2	2	2	3	3
$x^{(2)}$	2	1	2	3	1	2
y	+	+	-	+	-	+

   - représenter la frontière de décision d'un algo de 1-ppv
   - représenter la frontière de décision d'un algo de 3-ppv
   - combien d'erreurs ces deux classifieurs font-ils sur le jeu d'entrainement ?
   - quel sera la classe prédicte par ces deux classifieurs pour le point $(4; 0,5)$
2. étant donné n observations en p dimensions, quelle est l'erreur d'entrainement d'un 1-ppv (en classification) ?
3. entraîner un algo de k-ppv est-il plus long que de l'appliquer pour la prédiction de la classe d'une observation ?
4. Pour un algo de k-ppv, si on soupçonne que les données sont très bruitées, doit-on augmenter ou diminuer k ?

TD suite (extrait de "introduction au ML" de C. Azencott)

1. On cherche à prédire si une boisson est un thé ou un café à partir des données suivantes :

Volume (mL)	250	100	125	250
Caféine (g)	0,025	0,010	0,050	0,100
Boisson	Thé	Thé	Café	Café

   - En utilisant l'algo du 1-ppv, avec une distance euclidienne, quelle est l'étiquette prédite pour une boisson de 125mL, contenant 0,015g de caféine ?
   - cette prédiction ne semble pas très correcte, quel en est la cause et comment peut-on y rémedier ?

1 Introduction / Rappel

2 k-plus proches voisins (k-ppv) (IML Ch 8), (ESL Sec 13.3)

3 Classification naïve bayésienne (IML Ch 4, Sec 4.4)

Principe

En anglais

Naive Bayes Classifier

Principe

En anglais

Naive Bayes Classifier

Caractéristiques

- modélisation probabiliste du problème
- implique de maîtriser le théorème de Bayes

Principe

Caractéristiques

- modélisation probabiliste du problème
- implique de maîtriser le théorème de Bayes

Fonctionnement

- estimation de la probabilité de y en fonction de x
- utilisation d'une hypothèse simplificatrice (d'indépendance conditionnelle entre les variables)

Exemple introductif

Classification probabiliste de fruits

Labelliser un fruit comme "poire" (classe C_1) ou "pomme" (classe C_2)

1. utiliser la probabilité *a priori* $P(C_k)$ pour une décision
   Si $P(C_1) > P(C_2)$ alors c'est une poire

Exemple introductif

Classification probabiliste de fruits

Labelliser un fruit comme "poire" (classe C_1) ou "pomme" (classe C_2)

1. utiliser la probabilité *a priori* $P(C_k)$ pour une décision
   Si $P(C_1) > P(C_2)$ alors c'est une poire
   mais, cette probabilité est la même quelle que soit x, on prend toujours la même décision (avant même d'avoir pu observer x)

Exemple introductif

Classification probabiliste de fruits

Labelliser un fruit comme "poire" (classe C_1) ou "pomme" (classe C_2)

1. utiliser la probabilité *a priori* $P(C_k)$ pour une décision
   Si $P(C_1) > P(C_2)$ alors c'est une poire
2. utiliser les probabilités *a posteriori* $P(C_k|x)$
   Si $P(C_1|x) > P(C_2|x)$ alors c'est une poire

Exemple introductif

Classification probabiliste de fruits

Labelliser un fruit comme "poire" (classe C_1) ou "pomme" (classe C_2)

1. utiliser la probabilité *a priori* $P(C_k)$ pour une décision
   Si $P(C_1) > P(C_2)$ alors c'est une poire
2. utiliser les probabilités *a posteriori* $P(C_k|x)$
   Si $P(C_1|x) > P(C_2|x)$ alors c'est une poire
   mais, comment estimer ces quantités en pratique ?

Exemple introductif - suite

Avec un ensemble de caractéristiques (variables) $x \in \mathbb{R}^d$ pour décrire un fruit, on définit les probabilités conditionnelles :

$$P(x|C_1) \text{ et } P(x|C_2)$$

Exemple introductif - suite

Avec un ensemble de caractéristiques (variables) $x \in \mathbb{R}^d$ pour décrire un fruit, on définit les probabilités conditionnelles :

$$P(x|C_1) \text{ et } P(x|C_2)$$

On va mettre en œuvre le théorème de Bayes pour déterminer les probabilités *a posteriori* :

$$\forall k \in \{1,2\} \qquad P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_{l \in \{1,2\}} P(x|C_l)P(C_l)}$$

Exemple introductif - suite

Avec un ensemble de caractéristiques (variables) $x \in \mathbb{R}^d$ pour décrire un fruit, on définit les probabilités conditionnelles :

$$P(x|C_1) \text{ et } P(x|C_2)$$

On va mettre en œuvre le théorème de Bayes pour déterminer les probabilités *a posteriori* :

$$\forall k \in \{1,2\} \qquad P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_{l \in \{1,2\}} P(x|C_l)P(C_l)}$$

En d'autres termes, on affecte x à la classe dont la probabilité *a posteriori* est la plus forte.

Exemple introductif - suite

Avec un ensemble de caractéristiques (variables) $x \in \mathbb{R}^d$ pour décrire un fruit, on définit les probabilités conditionnelles :

$$P(x|C_1) \text{ et } P(x|C_2)$$

On va mettre en œuvre le théorème de Bayes pour déterminer les probabilités *a posteriori* :

$$\forall k \in \{1,2\} \qquad P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_{l \in \{1,2\}} P(x|C_l)P(C_l)}$$

En d'autres termes, on affecte x à la classe dont la probabilité *a posteriori* est la plus forte.
Mais comment estimer $P(x|C_k)$?

Notations

- Ensemble des classes $\{C_1, \cdots, C_K\}$ de probabilité *a priori* $P(C_k) = P(y = C_k) \quad \forall K$
- espace de caractéristiques (i.e. "features" en anglais) $\mathcal{X}$ (par ex. $\mathbb{R}^d$
- probabilité *a posteriori* $P(y = C_k|X = x) = P(C_k|x)$
- loi conditionnelle de x à la classe C_k $P(X = x|y = C_k) = P(x|C_k)$
- loi marginale de X $P_X(x) = P(X = x) = \sum_{k=1}^{K} P(x|C_k)P(C_k)$

Problème

Trouver la classe de x par une approche probabiliste grâce au Maximum A Posteriori (MAP) :

$$\begin{aligned} C_{\text{MAP}} &= \underset{k}{\text{argmax}}\, P(C_k|x) \\ &= \underset{k}{\text{argmax}}\, \frac{P(x|C_k)P(C_k)}{P(x)} \end{aligned}$$

Problème

Trouver la classe de x par une approche probabiliste grâce au Maximum A Posteriori (MAP) :

$$\begin{aligned} C_{\text{MAP}} &= \underset{k}{\text{argmax}}\, P(C_k|x) \\ &= \underset{k}{\text{argmax}}\, \frac{P(x|C_k)P(C_k)}{P(x)} \\ &\propto P(x|C_k)P(C_k) \end{aligned}$$

Problème

Trouver la classe de x par une approche probabiliste grâce au Maximum A Posteriori (MAP) :

$$\begin{aligned} C_{\text{MAP}} &= \underset{k}{\text{argmax}}\, P(C_k|x) \\ &= \underset{k}{\text{argmax}}\, \frac{P(x|C_k)P(C_k)}{P(x)} \\ &\propto P(x|C_k)P(C_k) \end{aligned}$$

En pratique

On cherche donc à estimer les probabilités $P(x|C_k)$ et $P(C_k)$ à partir de l'ensemble des observations.

Problème

Trouver la classe de x par une approche probabiliste grâce au Maximum A Posteriori (MAP) :

$$\begin{aligned} C_{\text{MAP}} &= \underset{k}{\text{argmax}}\, P(C_k|x) \\ &= \underset{k}{\text{argmax}}\, \frac{P(x|C_k)P(C_k)}{P(x)} \\ &\propto P(x|C_k)P(C_k) \end{aligned}$$

En pratique

On cherche donc à estimer les probabilités $P(x|C_k)$ et $P(C_k)$ à partir de l'ensemble des observations.
Si on connaissait parfaitement ces quantités, on aurait le classifieur optimal - **le classifieur de Bayes**-.

Estimation des probabilités *a priori*

Estimation de $P(C_k)$

$$\hat{P}(C_k) = \frac{n_k}{n}$$

Proportion [4] relative de la classe C_k dans l'ensemble d'apprentissage.

4. NB : $\sum_k \hat{P}(C_k) = 1$

Estimation des probabilités *a priori*

Estimation de $P(C_k)$

$$\hat{P}(C_k) = \frac{n_k}{n} = \frac{\text{nb d'obs. de } C_k}{\text{nb d'obs. total}}$$

Proportion [4] relative de la classe C_k dans l'ensemble d'apprentissage.

4. NB : $\sum_k \hat{P}(C_k) = 1$

Estimation des probabilités conditionnelles

$$P(x|C_k) = p(x^{(1)}, \cdots, x^{(d)}|C_k)$$

Estimation des probabilités conditionnelles

$$\begin{aligned} P(x|C_k) &= p(x^{(1)}, \cdots, x^{(d)}|C_k) \\ &= p(x^{(1)}|C_k)p(x^{(2)}, \cdots, x^{(d)}|C_k, x^{(1)}) \end{aligned}$$

Estimation des probabilités conditionnelles

$$
\begin{aligned}
P(x|C_k) &= p(x^{(1)}, \cdots, x^{(d)}|C_k) \\
&= p(x^{(1)}|C_k)p(x^{(2)}, \cdots, x^{(d)}|C_k, x^{(1)}) \\
&= p(x^{(1)}|C_k)p(x^{(2)}|C_k, x^{(1)})p(x^{(3)}, \cdots, x^{(d)}|C_k, x^{(1)}, x^{(2)}) \\
&= \cdots
\end{aligned}
$$

Estimation des probabilités conditionnelles

$$
\begin{aligned}
P(x|C_k) &= p(x^{(1)}, \cdots, x^{(d)}|C_k) \\
&= p(x^{(1)}|C_k)p(x^{(2)}, \cdots, x^{(d)}|C_k, x^{(1)}) \\
&= p(x^{(1)}|C_k)p(x^{(2)}|C_k, x^{(1)})p(x^{(3)}, \cdots, x^{(d)}|C_k, x^{(1)}, x^{(2)}) \\
&= \cdots
\end{aligned}
$$

Difficile à estimer sans hypothèse supplémentaire...

Estimation des probabilités conditionnelles - sous hypothèse d'indépendance

Hypothèse simplificatrice

Pour simplifier les calculs, on introduit une hypothèse naïve [5] et peu réaliste en pratique :

5. D'où le nom de la méthode

Estimation des probabilités conditionnelles - sous hypothèse d'indépendance

Hypothèse simplificatrice

Pour simplifier les calculs, on introduit une hypothèse naïve [5] et peu réaliste en pratique :

l'indépendance des attributs étant donné la classe

5. D'où le nom de la méthode

Estimation des probabilités conditionnelles - sous hypothèse d'indépendance

Hypothèse simplificatrice

Pour simplifier les calculs, on introduit une hypothèse naïve [5] et peu réaliste en pratique :

l'indépendance des attributs étant donné la classe

$$P(x|C_k) = p(x^{(1)}|C_k)p(x^{(2)}|C_k)\cdots p(x^{(d)}|C_k)$$
$$= \prod_{i=1}^{d} P(x^{(i)}|C_k)$$

5. D'où le nom de la méthode

Estimation des probabilités conditionnelles - sous hypothèse d'indépendance

Hypothèse simplificatrice

Pour simplifier les calculs, on introduit une hypothèse naïve [5] et peu réaliste en pratique :

l'indépendance des attributs étant donné la classe

$$\hat{P}(x|C_k) = \prod_{i=1}^{d} \hat{P}(x^{(i)}|C_k)$$

Il ne nous reste plus qu'à obtenir $\hat{P}(x_i|C_k)$

5. D'où le nom de la méthode

Exemple

Détection de l'évasion fiscale

Id	Refund	Marital status	Tax Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

On est à la recherche de $P(\text{Evade}|x)$

Retour sur l'estimation de probabilité à partir des données

Le cas discret

$$\hat{P}(X^{(1)} = x^{(1)}|C_k) = \frac{|x_k^1|}{n_k}$$

Retour sur l'estimation de probabilité à partir des données

Le cas discret

$$\hat{P}(X^{(1)} = x^{(1)}|C_k) = \frac{|x_k^1|}{n_k} = \frac{\text{nb d'obs avec } x^i \text{ parmi } C_k}{\text{nb d'obs. de } C_k}$$

Retour sur l'estimation de probabilité à partir des données

Le cas discret

$$\hat{P}(X^{(1)} = x^{(1)}|C_k) = \frac{|x_k^1|}{n_k} = \frac{\text{nb d'obs avec } x^i \text{ parmi } C_k}{\text{nb d'obs. de } C_k}$$

Exemples

- $\hat{P}(\text{Status} = \text{Married}|\text{Evade} = \text{No}) = \ ?$
- $\hat{P}(\text{Status} = \text{Married}|\text{Evade} = \text{Yes}) = \ ?$

Retour sur l'estimation de probabilité à partir des données

Le cas discret

$$\hat{P}(X^{(1)} = x^{(1)}|C_k) = \frac{|x_k^1|}{n_k} = \frac{\text{nb d'obs avec } x^i \text{ parmi } C_k}{\text{nb d'obs. de } C_k}$$

Exemples

- $\hat{P}(\text{Status} = \text{Married}|\text{Evade} = \text{No}) = \frac{4}{7}$
- $\hat{P}(\text{Status} = \text{Married}|\text{Evade} = \text{Yes}) = 0$

Retour à l'exemple

Id	Refund	Marital status	Tax Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Retour à l'exemple

Id	Refund	Marital status	Tax Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Retour à l'exemple

Id	Refund	Marital status	Tax Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $\hat{P}(\text{Status} = \text{Married}|\text{Evade} = \text{No}) = \frac{4}{7}$
- $\hat{P}(\text{Status} = \text{Married}|\text{Evade} = \text{Yes}) = 0$

Retour sur l'estimation de probabilité à partir des données

Le cas continu

- approximation de la densité par un histogramme
  - discrétisation de l'intervalle en paquets
  - estimation de la probabilité d'apparition d'une valeur dans le paquet

Retour sur l'estimation de probabilité à partir des données

Le cas continu

- approximation de la densité par un histogramme
  - discrétisation de l'intervalle en paquets
  - estimation de la probabilité d'apparition d'une valeur dans le paquet
- estimation (paramétrique) de la densité de probabilité
  - hypothèse gaussienne (loi normale $\mathcal{N}(\mu_k, \sigma_k)$) sur la distribution de l'attribut
  - estimation de μ_k et σ_k à partir des données
  - $\hat{P}(X^{(i)} = x^{(i)}|C_k) = \mathcal{N}(x^{(i)}; \mu_k^i, \sigma_k^i)$

Retour sur l'estimation de probabilité à partir des données

Le cas continu

- estimation (paramétrique) de la densité de probabilité
  - hypothèse gaussienne (loi normale $\mathcal{N}(\mu_k, \sigma_k)$) sur la distribution de l'attribut
  - estimation de μ_k et σ_k à partir des données
  - $\hat{P}(X^{(i)} = x^{(i)}|C_k) = \mathcal{N}(x^{(i)}; \mu_k^i, \sigma_k^i)$

Exemples

- $\hat{P}(\text{Income}|\text{Evade} = \text{No})$ - $\mu = ?$, $\sigma = ?$
- $\hat{P}(\text{Income}|\text{Evade} = \text{Yes})$ - $\mu = ?$, $\sigma = ?$

Retour à l'exemple

Id	Refund	Marital status	Tax Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $\hat{P}(\text{Income}|\text{Evade} = \text{No})$ $\quad \mu = \frac{770}{7} = 110,\ \sigma^2 = 2975$
- $\hat{P}(\text{Income}|\text{Evade} = \text{Yes})$ $\quad \mu = 90,\ \sigma^2 = 25$

Exemple complet (simplifié)

Id	Refund	Marital status	Evade
1	Yes	Single	No
2	No	Married	No
3	No	Single	No
4	Yes	Married	No
5	No	Divorced	Yes
6	No	Married	No
7	Yes	Divorced	No
8	No	Single	Yes
9	No	Married	No
10	No	Single	Yes

Exemple complet (simplifié)

Probabilité *a priori*

- $\hat{P}(\text{Evade} = \text{No}) =$
- $\hat{P}(\text{Evade} = \text{Yes}) =$

Probabilité conditionnelle x^1

- $\hat{P}(\text{Refund} = \text{Yes}|\text{No}) =$
- $\hat{P}(\text{Refund} = \text{No}|\text{No}) =$
- $\hat{P}(\text{Refund} = \text{Yes}|\text{Yes}) =$
- $\hat{P}(\text{Refund} = \text{No}|\text{Yes}) =$

Probabilité conditionnelle x^2

- $\hat{P}(\text{Status} = \text{Single}|\text{No}) =$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{No}) =$
- $\hat{P}(\text{Status} = \text{Married}|\text{No}) =$
- $\hat{P}(\text{Status} = \text{Single}|\text{Yes}) =$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{Yes}) =$
- $\hat{P}(\text{Status} = \text{Married}|\text{Yes}) =$

Exemple complet (simplifié)

Probabilité *a priori*

- $\hat{P}(\text{Evade} = \text{No}) = \frac{7}{10}$
- $\hat{P}(\text{Evade} = \text{Yes}) = \frac{3}{10}$

Probabilité conditionnelle x^1

- $\hat{P}(\text{Refund} = \text{Yes}|\text{No}) =$
- $\hat{P}(\text{Refund} = \text{No}|\text{No}) =$
- $\hat{P}(\text{Refund} = \text{Yes}|\text{Yes}) =$
- $\hat{P}(\text{Refund} = \text{No}|\text{Yes}) =$

Probabilité conditionnelle x^2

- $\hat{P}(\text{Status} = \text{Single}|\text{No}) =$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{No}) =$
- $\hat{P}(\text{Status} = \text{Married}|\text{No}) =$
- $\hat{P}(\text{Status} = \text{Single}|\text{Yes}) =$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{Yes}) =$
- $\hat{P}(\text{Status} = \text{Married}|\text{Yes}) =$

Exemple complet (simplifié)

Probabilité *a priori*

- $\hat{P}(\text{Evade} = \text{No}) = \frac{7}{10}$
- $\hat{P}(\text{Evade} = \text{Yes}) = \frac{3}{10}$

Probabilité conditionnelle x^1

- $\hat{P}(\text{Refund} = \text{Yes}|\text{No}) = \frac{3}{7}$
- $\hat{P}(\text{Refund} = \text{No}|\text{No}) = \frac{4}{7}$
- $\hat{P}(\text{Refund} = \text{Yes}|\text{Yes}) = 0$
- $\hat{P}(\text{Refund} = \text{No}|\text{Yes}) = 1$

Probabilité conditionnelle x^2

- $\hat{P}(\text{Status} = \text{Single}|\text{No}) =$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{No}) =$
- $\hat{P}(\text{Status} = \text{Married}|\text{No}) =$
- $\hat{P}(\text{Status} = \text{Single}|\text{Yes}) =$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{Yes}) =$
- $\hat{P}(\text{Status} = \text{Married}|\text{Yes}) =$

Exemple complet (simplifié)

Probabilité *a priori*

- $\hat{P}(\text{Evade} = \text{No}) = \frac{7}{10}$
- $\hat{P}(\text{Evade} = \text{Yes}) = \frac{3}{10}$

Probabilité conditionnelle x^1

- $\hat{P}(\text{Refund} = \text{Yes}|\text{No}) = \frac{3}{7}$
- $\hat{P}(\text{Refund} = \text{No}|\text{No}) = \frac{4}{7}$
- $\hat{P}(\text{Refund} = \text{Yes}|\text{Yes}) = 0$
- $\hat{P}(\text{Refund} = \text{No}|\text{Yes}) = 1$

Probabilité conditionnelle x^2

- $\hat{P}(\text{Status} = \text{Single}|\text{No}) = \frac{2}{7}$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{No}) = \frac{1}{7}$
- $\hat{P}(\text{Status} = \text{Married}|\text{No}) = \frac{4}{7}$
- $\hat{P}(\text{Status} = \text{Single}|\text{Yes}) = \frac{2}{3}$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{Yes}) = \frac{1}{3}$
- $\hat{P}(\text{Status} = \text{Married}|\text{Yes}) = 0$

Exemple complet (simplifié)

Probabilité *a priori*

- $\hat{P}(\text{Evade} = \text{No}) = \frac{7}{10}$
- $\hat{P}(\text{Evade} = \text{Yes}) = \frac{3}{10}$

Probabilité conditionnelle x^1

- $\hat{P}(\text{Refund} = \text{Yes}|\text{No}) = \frac{3}{7}$
- $\hat{P}(\text{Refund} = \text{No}|\text{No}) = \frac{4}{7}$
- $\hat{P}(\text{Refund} = \text{Yes}|\text{Yes}) = 0$
- $\hat{P}(\text{Refund} = \text{No}|\text{Yes}) = 1$

Probabilité conditionnelle x^2

- $\hat{P}(\text{Status} = \text{Single}|\text{No}) = \frac{2}{7}$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{No}) = \frac{1}{7}$
- $\hat{P}(\text{Status} = \text{Married}|\text{No}) = \frac{4}{7}$
- $\hat{P}(\text{Status} = \text{Single}|\text{Yes}) = \frac{2}{3}$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{Yes}) = \frac{1}{3}$
- $\hat{P}(\text{Status} = \text{Married}|\text{Yes}) = 0$

Prédiction pour $x = (\text{Refund} = \text{No}, \text{Status} = \text{Married})$

Exemple complet (simplifié)

Probabilité *a priori*

- $\hat{P}(\text{Evade} = \text{No}) = \frac{7}{10}$
- $\hat{P}(\text{Evade} = \text{Yes}) = \frac{3}{10}$

Probabilité conditionnelle x^1

- $\hat{P}(\text{Refund} = \text{Yes}|\text{No}) = \frac{3}{7}$
- $\hat{P}(\textbf{Refund} = \textbf{No}|\textbf{No}) = \frac{4}{7}$
- $\hat{P}(\text{Refund} = \text{Yes}|\text{Yes}) = 0$
- $\hat{P}(\textbf{Refund} = \textbf{No}|\textbf{Yes}) = 1$

Probabilité conditionnelle x^2

- $\hat{P}(\text{Status} = \text{Single}|\text{No}) = \frac{2}{7}$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{No}) = \frac{1}{7}$
- $\hat{P}(\textbf{Status} = \textbf{Married}|\textbf{No}) = \frac{4}{7}$
- $\hat{P}(\text{Status} = \text{Single}|\text{Yes}) = \frac{2}{3}$
- $\hat{P}(\text{Status} = \text{Divorced}|\text{Yes}) = \frac{1}{3}$
- $\hat{P}(\textbf{Status} = \textbf{Married}|\textbf{Yes}) = 0$

Prédiction pour $x = (\textbf{Refund} = \textbf{No}, \textbf{Status} = \textbf{Married})$

Exemple complet (simplifié) - prédiction

$$\text{Prédiction pour } x = (\text{Refund} = \text{No}, \text{Status} = \text{Married})$$

$$\begin{aligned}\hat{P}(x|\text{No}) &= \hat{P}(\text{Refund} = \text{No}|\text{No}) \times \hat{P}(\text{Status} = \text{Married}|\text{No}) \\ &= \frac{4}{7} \times \frac{4}{7}\end{aligned}$$

$$\begin{aligned}\hat{P}(x|\text{Yes}) &= \hat{P}(\text{Refund} = \text{No}|\text{Yes}) \times \hat{P}(\text{Status} = \text{Married}|\text{Yes}) \\ &= 1 \times 0\end{aligned}$$

Exemple complet (simplifié) - prédiction

$$\text{Prédiction pour } x = (\text{Refund} = \text{No}, \text{Status} = \text{Married})$$

$$\hat{P}(x|\text{No}) = \hat{P}(\text{Refund} = \text{No}|\text{No}) \times \hat{P}(\text{Status} = \text{Married}|\text{No})$$

$$\hat{P}(x|\text{Yes}) = \hat{P}(\text{Refund} = \text{No}|\text{Yes}) \times \hat{P}(\text{Status} = \text{Married}|\text{Yes})$$

Décision

$$\hat{P}(x|\text{No})P(\text{No}) > \hat{P}(x|\text{Yes})P(\text{Yes})$$
$$\Leftrightarrow \hat{P}(\text{No}|x) > \hat{P}(\text{Yes}|x)$$

On décide/prédit donc que x appartient à la classe "Evade=No".

Limites de la méthode

- Quand la valeur d'un attribut n'a pas été observée, sa probabilité conditionnelle est 0 et alors toute l'expression (à cause du produit) est nulle.
  Cela peut arriver quand on a peu de données [6] (ou beaucoup de modalités).
- L'hypothèse d'indépendance ne permet (logiquement) pas de prendre en compte la redondance d'information (correlation) entre des attributs
  à l'extrême, une information dupliquée verra son "poids" mis au carré dans la décision ...

6. Dans ce cas, on utilisera d'autres estimateurs (Laplace, M-estimateur)...

Bilan

- algorithme simple et souvent efficace en pratique (baseline)
- prédiction en temps constant pour toute nouvelle observation
- méthode probabiliste dont les décisions sont (relativement) interprétables
- hypothèse forte sur les données d'indépendance des variables (mais qui fonctionne assez bien en pratique)
- sensible au bruit dans les données et aux modalités non observées
- hypothèse supplémentaires à prendre en compte pour gérer des variables continues
- robustes aux valeurs manquantes

Exemple classique

Spam filter

Soit un ensemble de mots $\{w^{(1)}, \cdots, w^{(d)}\}$ dont on estime la fréquence d'apparition dans des courriels légitimes ou frauduleux ("Ham or Spam").
Pour un nouveau courriel M, on s'intéresse alors à :

$$\hat{P}(M|C) = \prod_{i=1}^{d} \hat{P}(w^{(i)}|C)$$

Exemple classique

Spam filter

Soit un ensemble de mots $\{w^{(1)}, \cdots, w^{(d)}\}$ dont on estime la fréquence d'apparition dans des courriels légitimes ou frauduleux ("Ham or Spam").
Pour un nouveau courriel M, on s'intéresse alors à :

$$\hat{P}(M|C) = \prod_{i=1}^{d} \hat{P}(w^{(i)}|C)$$

Jeu du chat et de la souris

Les spammer vont alors tenter d'utiliser des mots avec une probabilité $\hat{P}(w^i|\text{ham})$ forte ou alors avec une probabilité $\hat{P}(w^i|\text{spam})$ très faible jusqu'à présent.

Exemple classique

Spam filter

Soit un ensemble de mots $\{w^{(1)}, \cdots, w^{(d)}\}$ dont on estime la fréquence d'apparition dans des courriels légitimes ou frauduleux ("Ham or Spam").
Pour un nouveau courriel M, on s'intéresse alors à :

$$\hat{P}(M|C) = \prod_{i=1}^{d} \hat{P}(w^{(i)}|C)$$

Jeu du chat et de la souris

Le filtre doit alors mettre à jour ses estimations pour s'adapter à ces nouvelles pratiques.

Exemple classique

Spam filter

Soit un ensemble de mots $\{w^{(1)}, \cdots, w^{(d)}\}$ dont on estime la fréquence d'apparition dans des courriels légitimes ou frauduleux ("Ham or Spam").
Pour un nouveau courriel M, on s'intéresse alors à :

$$\hat{P}(M|C) = \prod_{i=1}^{d} \hat{P}(w^{(i)}|C)$$

Jeu du chat et de la souris

Les spammer vont alors tenter d'utiliser des mots avec une probabilité $\hat{P}(w^i|\text{ham})$ forte ou alors avec une probabilité $\hat{P}(w^i|\text{spam})$ très faible jusqu'à présent.

Exemple classique

Spam filter

Soit un ensemble de mots $\{w^{(1)}, \cdots, w^{(d)}\}$ dont on estime la fréquence d'apparition dans des courriels légitimes ou frauduleux ("Ham or Spam").
Pour un nouveau courriel M, on s'intéresse alors à :

$$\hat{P}(M|C) = \prod_{i=1}^{d} \hat{P}(w^{(i)}|C)$$

Jeu du chat et de la souris

Le filtre doit alors mettre à jour ses estimations pour s'adapter à ces nouvelles pratiques.

Exemple classique

Spam filter

Soit un ensemble de mots $\{w^{(1)}, \cdots, w^{(d)}\}$ dont on estime la fréquence d'apparition dans des courriels légitimes ou frauduleux ("Ham or Spam").
Pour un nouveau courriel M, on s'intéresse alors à :

$$\hat{P}(M|C) = \prod_{i=1}^{d} \hat{P}(w^{(i)}|C)$$

Jeu du chat et de la souris

etc ...

TD - exercice 1 - application à des données qualitatives

cf feuille de TD