Université de Montréal

Département d'informatique et de recherche opérationnelle

IFT3390/6390
Fondements de l'apprentissage machine

http://www.iro.umontreal.ca/~vincentp/ift3390

Cinquième cours:

Régression multiple et classifieur multiclasse. Rappels de proba. Classifieur de Bayes. Classifieur de Bayes Naïf

Professeur: Pascal Vincent

Laboratoire d'Informatique des Systèmes d'Apprentissage

Au programme aujourd'hui

- Régression multiple et classification multicalsse
- Rappels de bases de probabilités.
- Classifieur de Bayes.
- Classifieur de Bayes Naïf.
- Comment obtenir un classifieur à partir de la probabilité jointe P(X,Y).

Régression multiple

- Régression multiple: lorsque la cible n'est pas un unique scalaire réel mais un *vecteur* de m réels.
- On peut construire m régresseurs ordinaires. (Ex: fenêtre de Parzen régression)
- On peut aussi adapter un algorithme pour manipuler des vecteurs plutôt que des scalaires.
- Ex, pour Parzen régression, ça revient au même: on peut le voir comme m régresseurs de Parzen qui font chacun la *moyenne pondérée de cibles scalaires*, ou comme *un* Parzen qui fait une *moyenne pondérée de vecteurs*.

Classifieur binaire (rappel)

- On a vu que pour la classification binaire (2 classes), on utilise souvent un classifieur qui calcule une sortie scalaire (une seule valeur réelle).

- La fonction de décision consiste alors à comparer la sortie avec un *seuil*: On décide la première classe si la *sortie* est inférieure au *seuil* et la deuxième si la sortie est supérieure.

- Un cas fréquent est d'utiliser un *seuil* de 0, lorsque la sortie est réelle, ce qui revient à ne considérer que son signe.

- Un autre cas fréquent, est d'utiliser un *seuil* de 0.5, lorsque la *sortie* se situe dans [0, 1] et qu'on l'interprète comme la probabilité d'une des classes sachant l'entrée x. Notez que la probabilité de l'autre classe est alors *1-sortie*. Un seuil de 0.5 permet ainsi de choisir la classe la plus probable.

Classification multiclasse

- Voici une autre approche qui peut s'appliquer au cas binaire, mais se généralise au cas où on a m classes: l'algo produit en sortie un vecteur de dimension m, qui va contenir un "score" pour chaque classe.
- La fonction de décision consiste alors à décider la classe qui a obtenu le score le plus élevé: $décision = \arg\max(sortie)$.
- Quand les éléments sont positifs ou nuls et somment à 1, on peut interpréter la sortie j comme la probabilité de la jéme classe, sachant l'entrée.
- Remarquez qu'avec deux classes, choisir la classe avec le score le plus élevé, revient à considérer le signe de la différence entre les deux scores.

Comment obtenir un classifieur avec un algo de régresssion

- Dans le cas binaire, on peut effectuer une régression avec des cibles $Y \in \{0, 1\}$
  Un algorithme de régression parfait donnerait une prédiction

$$f(x) = E[Y|X = x] = P(Y = 1|X = x)$$

- Dans le cas de classification multiclasse (Y indique le numéro d'une classe parmi m classes), on peut effectuer une régression multiple avec une cible codée en "un parmi plusieurs" (one-hot) qui génère pour un y donné un vecteur de dimension m dont tous les éléments sont à 0 sauf l'élément y (correspondant à la varie classe) qui vaut 1.
  Ex: pour coder la deuxième classe parmi 4: $\text{onehot}_4(2) = (0, 1, 0, 0)$
  Une régression multiclasse parfaite prédirait un vecteur

$$\begin{aligned} f(x) &= E[\text{onehot}_m(Y)|X = x] \\ &= (P(Y = 1|X = x), \ldots, P(Y = m|X = x)) \end{aligned}$$

- Ainsi on peut dériver les algos de Parzen classification (binaire et multiclasse) des algos de Parzen régression (scalaire et multiple).

Comment obtenir un classifieur avec un algo de régresssion

- Mais on ne sait pas réaliser de régression "parfaite".
- Plus particulièrement, la prédiction d'un régresseur pourrait pour certains algos être négative, ou > 1.
- Aussi ces probabilités de classe ainsi obtenues (les estimés de $P(Y = j|X = x)$) ne sont pas garantis de sommer à 1.
- Cela peut néanmoins donner un bon algo de classification.

Approche probabiliste de l'apprentissage

- On suppose que les données sont générées par un processus inconnu.
- X,Y est vu comme une paire de variables aléatoires, distribuées selon une loi de probabilité inconnue P(X,Y).
- X (une variable vectorielle) est elle-même vue comme un ensemble de variables aléatoires scalaires. $P(X,Y) = P(X_{[1]}, \ldots, X_{[d]}, Y)$

Rappels de proba

Voir les 9 premiers transparents de la revue de proba et stats de Sam Roweis.

- Variable aléatoire discrète et continue
- Probabilité jointe
- Probabilité marginale, marginalisation
- Probabilité conditonnelle
- Règle de Bayes
- Indépendance

Classifieur de Bayes

ou comment construire un classifieur à partir d'estimations de densité

- On sépare l'ensemble d'entraînement en m sous-ensembles contenant chacun tous les points d'une même classe.
- On entraîne un estimateur de densité sur chacun: $c \in \{1, .., m\}$

$$\hat{p}_c(x) \simeq P(X = x|Y = c)$$

- On détermine les probabilités à priori de chaque classe $\hat{P}_c = \frac{n_c}{n} \simeq P(Y = c)$
(par ex. en comptant leurs proportions relatives dans l'ensemble d'apprentissage)
- On applique la règle de Bayes pour obtenir la probabilité à postériori des classes au point x.
- On choisit la plus probable.

$$\overbrace{P(Y = c|X = x)}^{posterior} = \frac{\overbrace{P(X = x|Y = c)}^{\substack{class\text{-}conditional\ density \\ (likelihood)}}\overbrace{P(Y = c)}^{prior}}{P(X = x)}$$

$$= \frac{P(X = x|Y = c)P(Y = c)}{\sum_{c'=1}^{m} P(X = x|Y = c')P(Y = c')}$$

$$\simeq \frac{\hat{p}_c(x)\hat{P}_c}{\sum_{c'=1}^{m} \hat{p}_{c'}(x)\hat{P}_{c'}}$$

Classifieur de Bayes Naïf

- Dans le classifieur de Bayes Naïf, on suppose, pour chaque classe $c \in \{1, .., m\}$ que, étant donné c, les composantes de X sont indépendantes:

$$P(X = x|Y = c) = P(X_{[1]} = x_{[1]}|Y = c)P(X_{[2]} = x_{[2]}|Y = c)...P(X_{[d]} = x_{[d]}|Y = c)$$
$$\hat{p}_c(x) = \hat{p}_{c,1}(x_{[1]})\hat{p}_{c,2}(x_{[2]})...\hat{p}_{c,d}(x_{[d]})$$

- Il suffit donc de modéliser des densités univariées, les $\hat{p}_{c,j}(x_{[j]})$ ce qui est une tâche facile (univariée == dimension 1: pas de fléau de la dimensionalité; les méthodes de type histogramme ou Parzen fonctionnent plutôt bien).

- On construit ensuite un classifieur de Bayes à partir des estimateurs $\hat{p}_c(x)$ ainsi obtenus.

Autre façon de construire un classifieur à partir d'un bon estimateur de densité

- On estime la probabilité jointe P(X,Y)
- On peut calculer la probabilité conditionnelle de la classe c:

$$P(Y = c|X = x) = \frac{P(X = x, Y = c)}{P(X = x)}$$

$$= \frac{P(X = x, Y = c)}{\sum_{c'=1}^{m} P(X = x, Y = c')}$$

- Notez que les proba de classe (conditionelles à x) sont proportionnelles aux probas jointes. Le dénominateur est une simple normalisation pour qu'elles somment à 1.
- Cette technique est utilisable du moment qu'on n'a pas un nombre gigantesque de classes.