Fiche ML

Charles Vin

S2-2023

1 Généralité

- Fonction de perte : quantifie l'erreur associé à une décision. Erreur simple : A chaque fois qu'on se trompe, on compte 1 : 0-1 loss
- Risque : Proba de se tromper, $R(y_i|x) = \sum_j l(y_i, y_j)P(y_j|x)$

2 Arbre de décision

3 Classfieur bayesien

On a :

- $P(y)$ fréquence des classe dans le dataset
- $P(x|y)$ les points de notre jeux de donnée. Graphiquement : les points coloriés

On cherche :

$$\arg\max_y P(y|x) = \arg\max_y \frac{P(x|y)P(y)}{P(x)}.$$

$P(x)$ difficile à calculer = répartition des points dans l'espace, dans le graph 2d non colorié. En général très petit, uniquement utile pour générer des données, pas pour faire l'argmax (aka classifier). Classifier bayésien = le classifier qui minimise le risque = le meilleurs classifieur possible

4 Estimation de densité

4.1 Par histogramme

Définition 4.1 (Estimation par histogramme)**.**

- Cas discret : Comptage dans chaque classe puis normalisation par le nombre d'exemple N
- Cas continue : Discrétisation des valeurs puis comptage et normalisation

Importance de la discrétisation :

- Petit $\rightarrow$ sur-apprentissage,
- Trop grand $\rightarrow$ sous-apprentissage

Limite :

- Grande dimention $\rightarrow$ Perte de sens
- Effet de bord : petit changement dans les bins, gros changement d'estimation.

$\rightarrow$ Solution : Estimation par noyaux

4.2 Estimation de densité par noyaux

Intuition figure 1 : Plutôt que de décider d'une discrétisation a priori, l'estimation est faîte en centrant une fenêtre autour du point d'intérêt x_0 (dans un espace de dimension d). $\rightarrow$ Problème : pas continue (si on bouge la boite et qu'un point rentre dedans, ça fait faire un saut à la fonction)

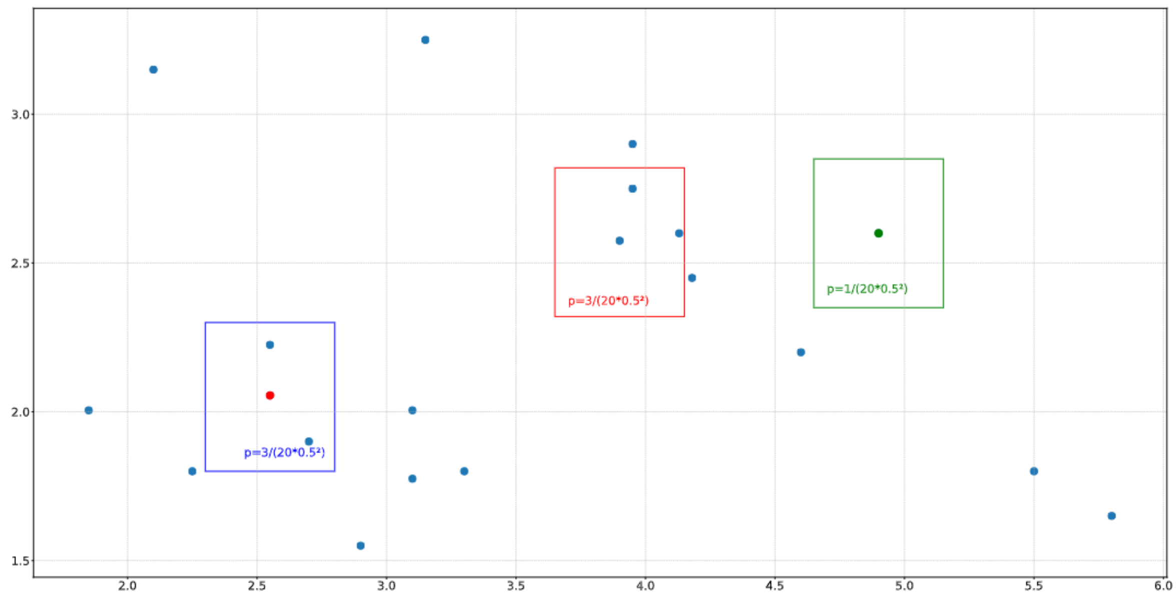

Figure 1 – Intuition de l'estimation par noyaux

4.2.1 Fenêtre de Parzen

On combine la solution précédente avec une densité/noyaux. Classiquement Gaussien. pour obtenir un truc lisse et continue

Définition 4.2 (Fenêtre de Parzen)**.** Soit $(x_1, \ldots, x_N) \sim f$ iid

$$\hat{f}_h(x) = \frac{1}{N * h} \sum_{i=1}^{N} K(\frac{x - x_i}{h}).$$

Avec K le noyaux **centrée et réduit sur** x , souvent une fonction gaussienne. Si c'est une fonction rectangle ça fonctionne aussi. Puis y'a plein d'autre noyaux possible.