Naïve Bayes classifiers

Data: $\vec{x}^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)})$ $\quad i=1,\ldots,N$

$y^{(i)} \in C_q \subset \{C_1, \ldots, C_Q\}$

Q classes.

Bayes' theorem:

$$P(Y=y_q \mid X=\vec{x}) = \frac{\overset{B}{P(X=\vec{x} \mid Y=y_q)}\ \overset{A}{P(Y=y_q)}}{\underset{C}{P(X=\vec{x})}}$$

A Prior: A.1 frequency of each class in the training set.

$$P(Y=y_q) = \frac{\sum_{i=1}^{N} I(y^{(i)}=y_q)}{N}$$

$$I(A) = \begin{cases} 1, & \text{if } A \text{ is true} \\ 0, & \text{otherwise}. \end{cases}$$

or A.2 use any additional information to inform your prior choice.

B Naive assumption:

$$P(X_1 = x_1, X_2 = x_2, \dots, X_p = x_p \mid Y = y_q) =$$

$$= \prod_{j=1}^{p} P(X_j = x_j \mid Y = y_q)$$

Wrong but it works!

B.1

$$P(X_j = x_j \mid Y = y_q) \approx \frac{\sum_{i=1}^{N} I(x_j^{(i)} = x_j \;\&\; y^{(i)} = y_q)}{\sum_{i=1}^{N} I(y^{(i)} = y_q)}$$

Laplace smoothing:

$$P(X_j = x_j \mid Y = y_q) = \frac{\sum_{i=1}^{N} \left[I(x_j^{(i)} = x_j \;\&\; y^{(i)} = y_q) \right] + 1}{\sum_{i=1}^{N} I(y^{(i)} = y_q) + N}$$

B.2 Assume a distribution and the construct estimators for the parameters of the distribution.

e.g. Gaussian:

$$P(X_j = x_j \mid Y = y_q) = \frac{1}{\sqrt{2\pi \sigma_{j,q}^2}} e^{\frac{-(x_j - \mu_{j,q})^2}{2\sigma_{j,q}}}$$

$$\mu_{j,q} = \frac{1}{N_q} \sum_{y^{(i)} \in C_q} x_j^{(i)}$$

sample mean for x_j.

$$\sigma_{j,q}^2 = \frac{1}{N_q - 1} \sum_{y^{(i)} \in C_q} \left(x_j^{(i)} - \mu_{j,q}\right)^2$$

Sample variance

c

$$P(X = \vec{x}) = \sum_{q=1}^{Q} \underbrace{P(X = \vec{x} \mid Y = y_q)}_{B} \underbrace{P(Y = y_q)}_{A}$$

$$= \sum_{q=1}^{Q} \left[A \times B \right]_q$$

$$P(Y = y_q \mid X = \vec{x}) = \frac{\left[A \times B \right]_q}{\sum_{q=1}^{Q} \left[A \times B \right]_q}$$

$$A_q = P(Y = y_q) = \frac{1}{N} \sum_{i=1}^{N} I(y^{(i)} = y_q)$$

$$B_q = \prod_{j=1}^{P} P(X_j = x_j \mid Y = y_q) = \prod_{j=1}^{P} \left[\frac{\sum_{i=1}^{N} I(x_j^{(i)} = x_j \; \& \; y^{(i)} = y_q)}{\sum_{i=1}^{N} I(y^{(i)} = y_q)} \right]$$

Outcome : Given $\vec{x}^{in}$

our NB classifier gives us a probability vector of $\vec{x}^{in}$ belonging to each of the Q classes

$$\underset{Q\times 1}{\vec{\pi}^{(NB)}} = \begin{bmatrix} \pi_1^{(NB)} \\ \vdots \\ \pi_Q^{(NB)} \end{bmatrix}$$

$$\pi_q^{(NB)} = \frac{[\text{A} \times \text{B}]_q}{\sum_{q=1}^{Q} [\text{A} \times \text{B}]_q}$$

$$\vec{1}^T \cdot \vec{\pi}^{(NB)} = 1 \quad \checkmark$$ (by construction!)

Classifier $\Rightarrow$ $\underbrace{\hat{y}^{(NB)} = \arg\max_q \vec{\pi}^{(NB)}}$

choose the maximum: MAP (maximum a posteriori)

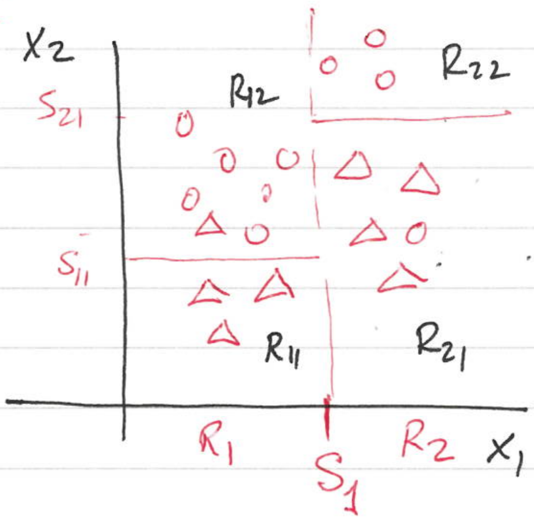
Leading to Random Forests

we start with Decision Trees

Discrete or continuous — Discrete or continuous

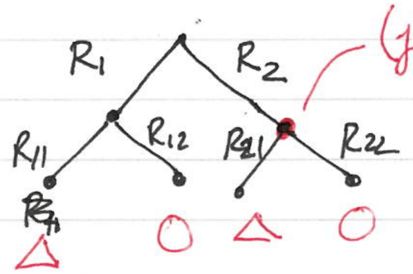
$$\vec{x} = (x_1, \dots, x_p) \longmapsto y$$

Cartoon:



Descriptor that is split

(j, s)



Threshold.

In this case: $(j=2,\ s=S_{21})$

Each split (j, s) corresponds to:
$$\begin{cases} \{x_j : x_j < s \equiv R_1(j,s)\} \\ \{x_j : x_j \geq s \equiv R_2(j,s)\} \end{cases}$$

Regression:

choose (j, s) that minimises a loss:

$$\min_{j,s} \left[\sum_{\vec{x}^{(i)} \in R_1(j,s)} \left(y^{(i)} - \bar{\bar{y}}_{R_1}\right)^2 + \sum_{\vec{x}^{(i)} \in R_2(j,s)} \left(y^{(i)} - \bar{\bar{y}}_{R_2}\right)^2 \right]$$

$$\bar{\bar{y}}_{R_1} = \frac{\sum_{i=1}^{N} I\left(x_j^{(i)} < s\right) \cdot y^{(i)}}{\sum_{i=1}^{N} I\left(x_j^{(i)} < s\right)}$$

[We represent each of the points in the region by the mean of the region]

- Optimise cost at every split: <u>greedily</u>.

- Continue until a stopping criterion is met:

e.g.
  - Plateau in optimisation
  - small # of points in region
  - using all predictors.

$$\boxed{\begin{array}{l} \text{Given } \vec{x}^{in} \\ \qquad \text{Find } R_{\vec{\alpha}} \;/\; \vec{x}^{in} \in R_{\vec{\alpha}} \\ \qquad\qquad \hat{y}^{DT} = \bar{y}_{R_{\vec{\alpha}}} \end{array}}$$

Mean of the region where $\vec{x}^{in}$ falls.