The classification problem

Given: TS $(\underline{x}_1, y_1) \ldots (\underline{x}_n\, y_n) \sim$ iid obs of $(\underline{X}, Y)$

$\underline{x}_i \in \mathbb{R}^p$

$y_i \in \{1 \ldots C\}$ categorical

Goal: Generate prediction rule $s(\underline{x})$ that predicts Y for given $\underline{x}$.

To measure performance of the rule we need a loss function $L(j, k)$ = loss incurred if $Y = j$ but we predict $s(\underline{x}) = k$

Always assume that $L(j, j) = 0$

Recall: In regression case (Y numerical) we commonly use $L(Y, s(\underline{x})) = (Y - s(\underline{x}))^2$

What is the optimal rule that minimizes the risk $R(s) = E_{\underline{X}, Y} L(Y, s(\underline{X}))$?

Define

$p_k(\underline{x})$ = conditional density of $\underline{X}$ given $Y = k$

$\pi_k = P(Y = k)$

$\Downarrow$

$$P(Y = k \mid \underline{X} = \underline{x}) = \pi_k\, p_k(x) \Big/ \sum_j \pi_j\, p_j(\underline{x})$$

$$R(s) = E_{\underline{X}, Y}(L(Y, s(\underline{X})))$$

$$= E_{\underline{x}}\, E_{Y|\underline{x}} \left(L(Y, s(\underline{x})) \right)$$

The minimization can be done pointwise for each $\underline{x}$

$$E_{Y|\underline{x}} \left(L(Y, s(\underline{x})) \right) = \sum_{k=1}^{C} p(Y=k\,|\,\underline{x}) \cdot L(k, s(\underline{x}))$$

$$s(\underline{x}) = \underset{i}{\operatorname{argmin}} \sum_{k=1}^{C} p(Y=k\,|\,\underline{x}) \cdot L(k, i)$$

Look at "unit loss" $L(k,i) = 1 - \delta(k,i)$

$$\delta(k,i) = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{otherwise} \end{cases}$$

For unit loss

$$s(\underline{x}) = \underset{i}{\operatorname{argmin}} \sum_{k=1}^{C} P(Y=k\,|\,\underline{x}) (1 - \delta(k,i))$$

$$= \underset{i}{\operatorname{argmax}} \sum_{k=1}^{C} P(Y=k\,|\,\underline{x})\, \delta(k,i)$$

$$= \underset{i}{\operatorname{argmax}}\; P(Y=i\,|\,\underline{x})$$

$\Rightarrow$ Predict the class that is most likely at $\underline{x}$ (This is Bayes rule for unit losses)

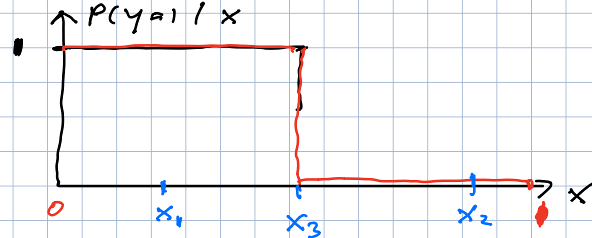
The optimal rule is called Bayes rule
The risk of the Bayes rule is called Bayes risk

In practice the conditional class

distributions $p(Y=k \mid \underline{X}=\underline{x})$ have to be estimated from the training sample.

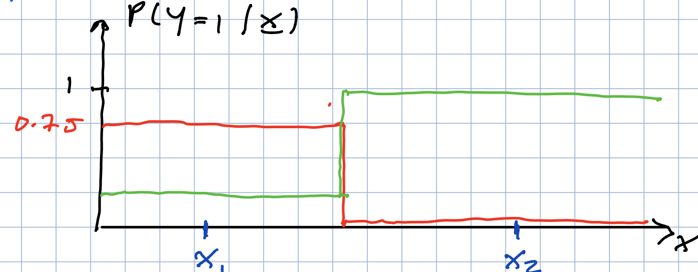
As long as $P(Y=k \mid \underline{X}=\underline{x})$ is $\neq 1$ for all k we cannot make perfect prediction for this articalar predictor value x.

Suppose $C=2$ and $P(Y=1 \mid \underline{x}) = P(Y=2 \mid \underline{x}) = 0.5$



$X \sim U[0,1]$

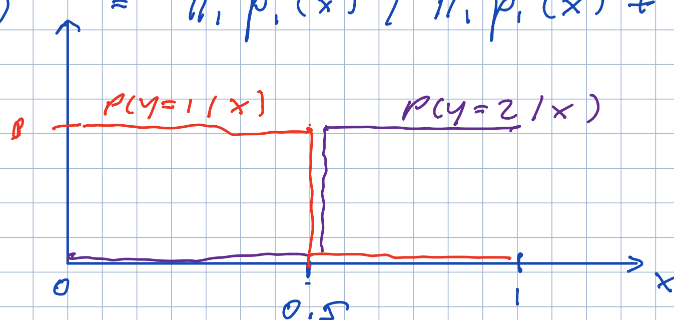
$y_1 = 1 \qquad y_2 = 2$



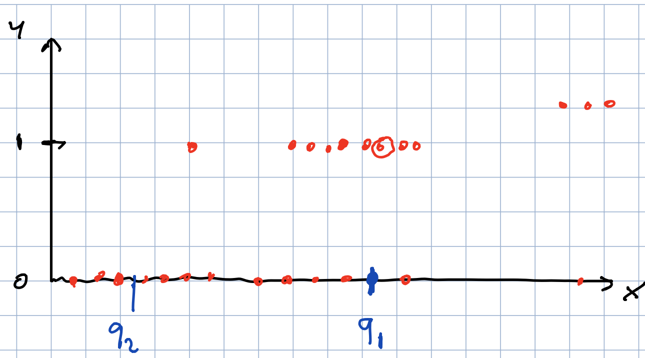
$p_1(x) \sim U[0, \tfrac{1}{2}] \qquad \pi_1 = 0.7$

$p_2(x) \sim U[\tfrac{1}{2}, 1] \qquad \pi_2 = 0.3$

$p(Y=1 \mid x) = \pi_1 p_1(x) / \pi_1 p_1(x) + \pi_2 p_2(x)$



For $x < 0.5$ $p_2(x) = 0$ $\Rightarrow$ $P(Y=1 | x) = 1$



... Training obs

$\hat{y}(q_1) = 1$

$\hat{y}(q_2) = 0$

$\hat{y}(q)$: find k-nearest tr