582631 Introduction to Machine Learning, Fall 2016
Exercise set IV
Model solutions

1.

(a) Using the assumption about class-conditional independence of X_1 and X_2, and denoting the density functions of conditional distributions as $f_{X|Y=c}$, we get from the Bayes formula:

$$\begin{aligned}
&P(Y=1|X_1=1,X_2=2)\\
&=\frac{f_{\mathbf{X}|Y=1}(1,2)P(Y=1)}{f_{\mathbf{X}|Y=1}(1,2)P(Y=1)+f_{\mathbf{X}|Y=-1}(1,2)P(Y=-1)}\\
&=\frac{f_{X_1|Y=1}(1)f_{X_2|Y=1}(2)}{f_{X_1|Y=1}(1)f_{X_2|Y=1}(2)+f_{X_1|Y=-1}(1)f_{X_2|Y=-1}(2)}\\
&=\frac{\frac{1}{\sigma_{+,1}}e^{-\frac{(1-\mu_{-,1})}{2\sigma^2_{+,1}}}\frac{1}{\sigma_{+,2}}e^{-\frac{(2-\mu_{-,2})^2}{2\sigma^2_{+,2}}}}{\frac{1}{\sigma_{+,1}}e^{-\frac{(1-\mu_{+,1})}{2\sigma^2_{+,1}}}\frac{1}{\sigma_{+,2}}e^{-\frac{(2-\mu_{+,2})^2}{2\sigma^2_{+,2}}}+\frac{1}{\sigma_{-,1}}e^{-\frac{(1-\mu_{-,1})}{2\sigma^2_{-,1}}}\frac{1}{\sigma_{-,2}}e^{-\frac{(2-\mu_{-,2})^2}{2\sigma^2_{-,2}}}}\\
&=\frac{\frac{1}{16}e^{-5/32}}{\frac{1}{16}e^{-5/32}+e^{-5/2}}\approx 0.3944.
\end{aligned}$$

(b) See Fig. 1.

(c) Both the naive Bayes classifier with Gaussian densities and QDA are based on the Bayes formula

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{p(\mathbf{x})},$$

and therefore, assuming equivalent class distributions $p(y)$, we only need to show that for the given class-conditional distribution implied by a naive Bayes classifier $p(\mathbf{x} \mid y)$, we can construct a QDA classifier with the same class-conditional distribution.

Since the naive Bayes model implies that X_1 and X_2 are independent given Y, we are looking for a bivariate Gaussian density for the QDA model that is equivalent to the product of two Gaussian densities. This tells us that the covariance matrix must be diagonal.

Furthermore, given that the class-conditional distribution under the QDA model must match that of the NB model, and in particular, that the means and variances of X_1 and X_2 must match those of the NB model, we arrive at the following QDA parameters:

$$\begin{aligned}
\boldsymbol{\mu}_+ &= (\mu_{+,1}, \mu_{+,2}) = (0,0),\\
\boldsymbol{\mu}_- &= (\mu_{-,1}, \mu_{-,2}) = (0,0),
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\Sigma}_+ &= \begin{bmatrix} \sigma^2_{+,1} & 0 \\ 0 & \sigma^2_{+,2} \end{bmatrix} = \begin{bmatrix} 16 & 0 \\ 0 & 16 \end{bmatrix},\\
\boldsymbol{\Sigma}_- &= \begin{bmatrix} \sigma^2_{-,1} & 0 \\ 0 & \sigma^2_{-,2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.
\end{aligned}$$

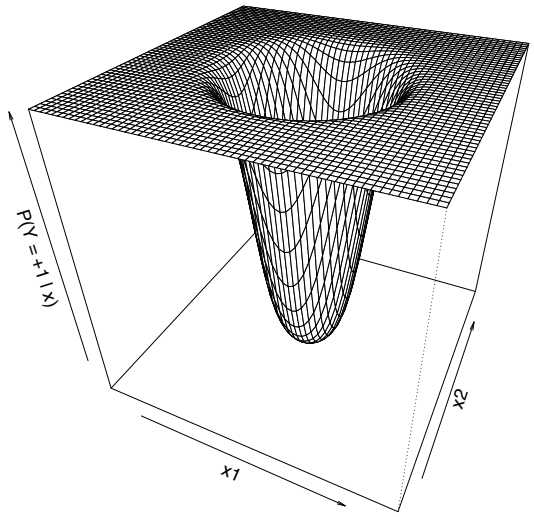

Figure 1: Illustration of the posterior probability $P(Y = +1 \,|\, \mathbf{x})$

Having made these choices, we can verify that the class-conditional distributions match. Fix $y \in \{-1, +1\}$ and consider a conditional density $p(\mathbf{x} \,|\, y)$ under the naive Bayes model.

$$\begin{aligned} p(\mathbf{x} \,|\, y) &= f_{\mathbf{X}|Y=y}(\mathbf{x}) \\ &= f_{X_1|Y=y}(x_1) f_{X_2|Y=y}(x_2) \\ &= \frac{1}{\sqrt{2\pi\sigma_{y,1}^2}} \exp\left(-\frac{(x_1 - \mu_{y,1})^2}{2\sigma_{y,1}^2}\right) \frac{1}{\sqrt{2\pi\sigma_{y,2}^2}} \exp\left(-\frac{(x_2 - \mu_{y,2})^2}{2\sigma_{y,2}^2}\right) \\ &= \frac{1}{2\pi\sigma_{y,1}\sigma_{y,2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{2} \frac{(x_i - \mu_{y,i})^2}{2\sigma_{y,i}^2}\right) \end{aligned}$$

We can now use Exercise 3.2a from last week to notice that this is a density function of bivariate normal distribution with expected value $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_y$ defined above:

$$\frac{1}{2\pi|\boldsymbol{\Sigma}_y|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_y^{-1} (\mathbf{x} - \boldsymbol{\mu}_y)\right).$$

Thus, the NB model is equivalent the the QDA model with the chosen parameters.

(d) Consider a classification problem with k classes. Both the Gaussian NB and the QDA classifiers require mean parameters for each class and each feature. The number of such parameters is kp.

For QDA, the number of free parameters required for the covariance matrices $\mathbf{\Sigma_1}, \ldots, \mathbf{\Sigma}_k$ is $\frac{kp(p+1)}{2}$. In the NB case, we only need the variance of each feature, so kp variance parameters in total.

In addition, both classifiers involve k class probabilities that must sum to one, which means that the required number of free parameters is $k-1$.

Thus, QDA requires altogether $(k-1)+kp+\frac{kp(p+1)}{2}$ parameters. NB requires $(k-1)+kp+kp$ parameters. The difference grows with p since the number of free paramaters is quadratic in p for QDA, but only linear in p for NB. Hence, for large p, QDA is more prone to overfitting and requires a larger sample size to reach its asymptotic error.

3. An example decision tree is shown In Figure 2. First the data is split into upper half R_1 and lower half $\{R_2, \ldots, R_6\}$, then the lower half is split into a left corner R_2 and the rest $\{R_3, \ldots, R_6\}$. The rest of the splits separate the right-most area from the remaining subset in the order R_3, R_4 and R_5.

The gains of the splits using misclassification error as impurity measure are tabulated in Table 1. The only split to have positive is the last. This is a consequence of the fact that the last split is the only one that leads to a change in the the majority class — the majority of the shots overall (in all of the data) and in each of the subsets $R_1, \ldots, R_5$ are goals, but within subset R_6 the saves make the majority.

In next week's exercises, you will find that the situation is somewhat different when the Gini index of the entropy is used to define impurity.

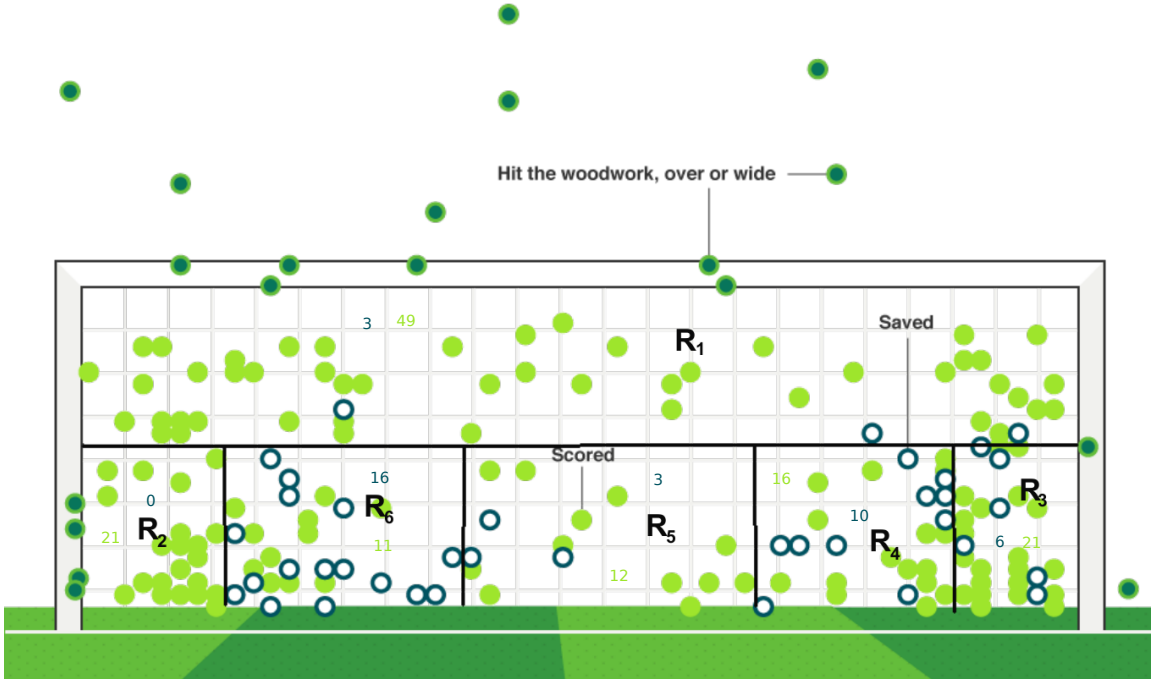

Figure 2: Penalties shot in World cup up to South Africa 2010

Table 1: Gains of the splits of the decision tree grown on the data set of Figure 2.

Split	$Q(D_1)$	$Q(D_2)$	$Q(D)$	gain(D_1, D_2)
R_1	3/52	35/116	38/168	0
R_2	0/21	35/95	35/116	0
R_3	6/27	29/68	35/95	0
R_4	10/26	19/42	29/68	0
R_5	3/15	11/27	19/42	0.119