Assignment 1
Machine Learning 10-701

ARJUN MENON
Carnegie Mellon University
October 1, 2013

1 Probability Review [Ahmed]

1.1 Why just 2 variables? Let's go for 3

1.1.1

From the law of conditional probability,

$$\begin{aligned}
\frac{\Pr(A, B \mid C)}{\Pr(B \mid C)} &= \frac{\Pr(A, B, C)}{\Pr(C)} \frac{\Pr(C)}{\Pr(B, C)} \\
&= \frac{\Pr(A, B, C)}{\Pr(B, C)} \\
&= \frac{\Pr(A, D)}{\Pr(D)} \\
&= \Pr(A \mid D) \\
&= \Pr(A \mid B, C) \qquad \square
\end{aligned}$$

1.1.2

$$\begin{aligned}
\sum_B \Pr(A, B \mid C) &= \sum_B \frac{\Pr(A, B, C)}{\Pr(C)} \\
&= \frac{\Pr(A, C)}{\Pr(C)} \\
&= \Pr(A \mid C) \qquad \square
\end{aligned}$$

1.1.3

Using result from Problem 1.1.1 and Problem 1.1.2,

$$\begin{aligned}
\sum_B \Pr(A \mid B, C) \Pr(B \mid C) &= \sum_B \Pr(A, B \mid C) \\
&= \Pr(A \mid C) \qquad \square
\end{aligned}$$

1.2 Evaluating Test Results

1.2.1

The probability that a transation succeeds given that it was handled by $A2$ is

$$\begin{aligned}
\Pr(Success \mid A = 2) &= \frac{\Pr(Success, A = 2)}{\Pr(A = 2)} \\
&= \frac{|Success, A = 2|}{|A = 2|} \\
&= \frac{2150}{2150 + 500} \\
&= 0.811
\end{aligned}$$

1.2.2

If we recommend $A2$ then we need to see that

$$\begin{aligned}
\Pr(Success \mid A = 2) &\geq \Pr(Success \mid A = 1) \\
0.811 &\geq \frac{6000}{6000 + 1700} \\
0.811 &\geq 0.779 \qquad \square
\end{aligned}$$

1.2.3

The statement about the probability of $A2$ handling a transaction successfully given $A1$ handled it successfully is given by,

$$\begin{aligned}
\Pr(A2_{suc} \mid A1_{suc}) &= \frac{\Pr(A2_{suc}, A1_{suc})}{\Pr(A1_{suc})} \\
&\geq \frac{\Pr(A2_{suc}) + \Pr(A1_{suc}) - 1}{\Pr(A1_{suc})} \\
&= \frac{\frac{2150}{2150+500} + \frac{6000}{6000+1700} - 1}{\frac{6000}{6000+1700}} \\
&= 0.757 \\
\Pr(A2_{suc} \mid A1_{suc}) &\geq 0.757 \qquad \square
\end{aligned}$$

1.3 Monty Hall Problem

Solving for $\Pr(car3 \mid open2, choose1)$ we get,

$$\begin{aligned}
\Pr(car3 \mid open2, choose1) &= \frac{\Pr(car3, open2 \mid choose1)}{\Pr(open2 \mid choose1)} \\
&= \frac{\Pr(open2 \mid choose1, car3) \Pr(car3 \mid choose1)}{\Pr(open2 \mid choose1)} \\
&= \frac{1 \times \frac{1}{3}}{\frac{1}{2}} \\
&= \frac{2}{3}
\end{aligned}$$

and solving for $\Pr(car1 \mid open2, choose1)$ we get,

$$\begin{aligned}
\Pr(car1 \mid open2, choose1) &= \frac{\Pr(car1, open2 \mid choose1)}{\Pr(open2 \mid choose1)} \\
&= \frac{\Pr(open2 \mid choose1, car1) \Pr(car1 \mid choose1)}{\Pr(open2 \mid choose1)} \\
&= \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} \\
&= \frac{1}{3}
\end{aligned}$$

which gives us,

$$\begin{aligned}
\frac{\Pr(car3 \mid open2, choose1)}{\Pr(car1 \mid open2, choose1)} &= \frac{\frac{2}{3}}{\frac{1}{3}} \\
&= 2 \qquad \square
\end{aligned}$$

2 Regression [Leila]

2.1 Linear Regression

Starting with the definition of the MLE estimate,

$$\begin{aligned}
\beta_{MLE} &= \arg\max_{\beta \in B} \Pr\left(Data \mid \beta\right) \\
&= \arg\max_{\beta \in B} \prod_{i=1}^{n} \Pr\left(y_i \mid x_i, \beta\right) \\
&= \arg\max_{\beta \in B} \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left(\frac{-1}{2\sigma^2}(y_i - \beta x_i)^2\right) \\
&= \arg\max_{\beta \in B} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2\right) \\
&= \arg\max_{\beta \in B} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\frac{-1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\right)
\end{aligned}$$

The expressions above is maximized when $(Y - X\beta)^T(Y - X\beta)$ is minimized where,

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$
$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

Minimizing $L = (Y - X\beta)^T(Y - X\beta)$ by taking the gradient with respect to β, setting to zero, and solving for β gives us the β_{MLE} as follows,

$$\begin{aligned}
L &= (Y - X\beta)^T(Y - X\beta) \\
L &= Y^TY - 2Y^TX\beta + \beta^TX^TX\beta \\
\nabla_\beta L &= 0 - 2Y^TX + (X^TX + (X^TX)^T)\beta \\
0 &= -2Y^TX + 2X^TX\beta \\
X^TX\beta &= X^TY \\
\beta_{MLE} &= (X^TX)^{-1}X^TY
\end{aligned}$$

□

2.2 Ridge Regression

Starting with the definition of the MAP estimate,

$$\begin{aligned}
\beta_{MAP} &= \arg\max_{\beta \in B} \Pr\left(\beta \mid Data\right) \\
&= \arg\max_{\beta \in B} \frac{\Pr\left(Data \mid \beta\right)\Pr\left(\beta\right)}{\Pr\left(Data\right)} \\
&\propto \arg\max_{\beta \in B} \Pr\left(Data \mid \beta\right)\Pr\left(\beta\right) \\
&= \arg\max_{\beta \in B} \left[\prod_{i=1}^{n} \Pr\left(y_i \mid x_i, \beta\right)\right] \Pr\left(\beta\right) \\
&= \arg\max_{\beta \in B} \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(y_i - \beta x_i)^2\right)\right] \exp\left(\frac{-\beta^T\beta}{2\lambda^2}\right) \\
&\propto \arg\max_{\beta \in B} \left[\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2\right)\right] \exp\left(\frac{-\beta^T\beta}{2\lambda^2}\right) \\
&= \arg\max_{\beta \in B} \left[\exp\left(\frac{-1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\right)\right] \exp\left(\frac{-\beta^T\beta}{2\lambda^2}\right) \\
&= \arg\max_{\beta \in B} \exp\left(\frac{-1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) + \frac{-\beta^T\beta}{2\lambda^2}\right)
\end{aligned}$$

The expressions above is maximized when $\frac{-1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) - \frac{\beta^T\beta}{2\lambda^2}$ is minimized where,

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$
$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

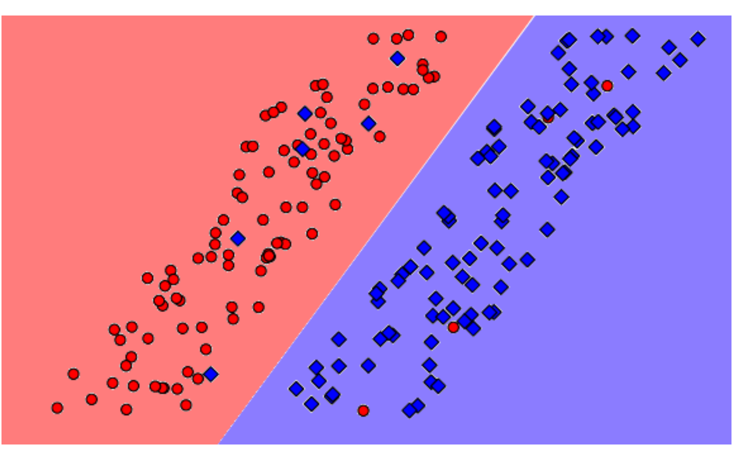
Since $\sigma = 1$, we can also get rid of the 2's and minimize $L = (Y - X\beta)^T(Y - X\beta) - \frac{\beta^T\beta}{\lambda^2}$ by taking the gradient with respect to β, setting to zero, and solving for β gives us the β_{MAP} as follows,

$$\begin{aligned}
L &= (Y - X\beta)^T(Y - X\beta) - \frac{\beta^T\beta}{\lambda^2} \\
L &= Y^TY - 2Y^TX\beta + \beta^TX^TX\beta - \lambda'\beta^T\beta \\
\nabla_\beta L &= 0 - 2Y^TX + (X^TX + (X^TX)^T)\beta - 2\lambda'\beta \\
0 &= -2Y^TX + 2X^TX\beta - 2\lambda'\beta \\
(X^TX + \lambda')\beta &= X^TY \\
\beta_{MAP} &= (X^TX + \lambda')^{-1}X^TY
\end{aligned}$$

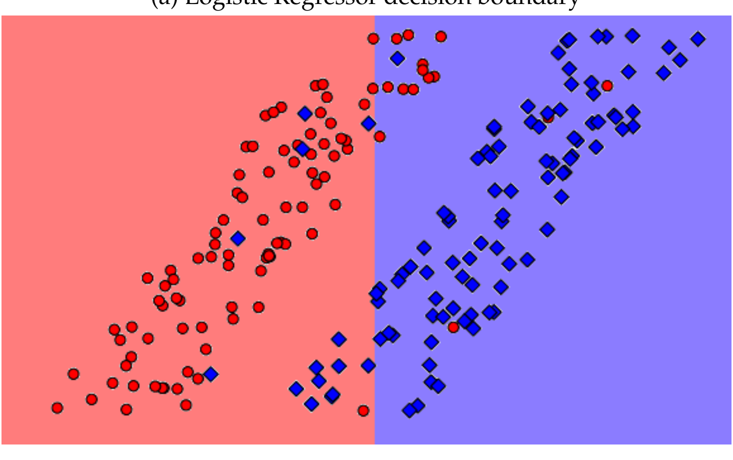
□

3 Classification [Dougal]

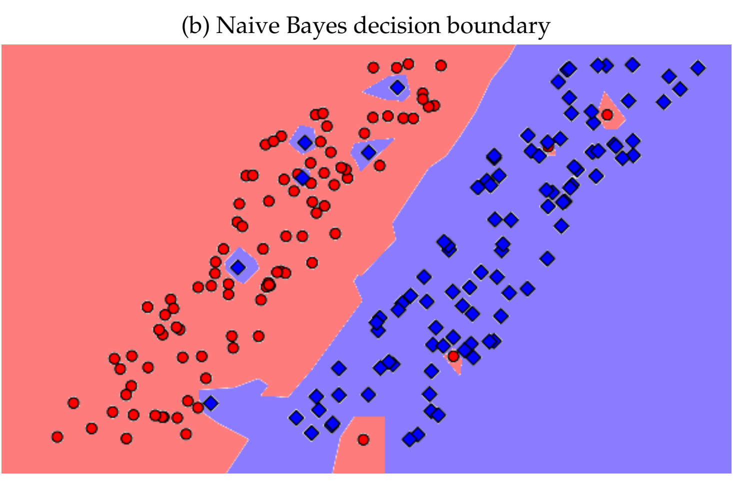
3.1 Drawing decision boundaries



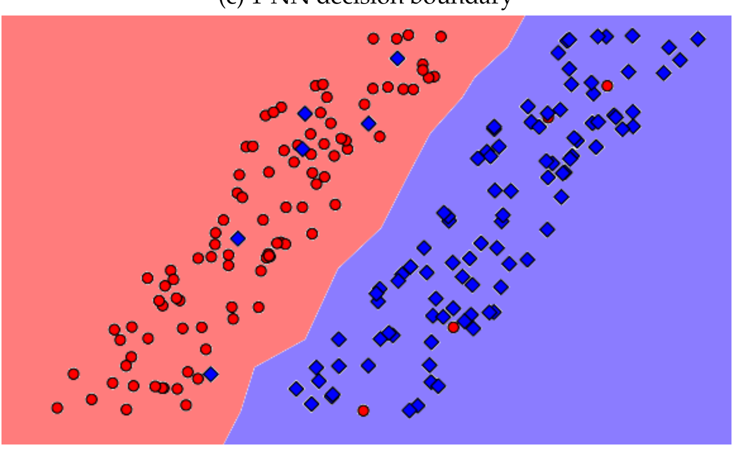
(a) Logistic Regressor decision boundary



(b) Naive Bayes decision boundary



(c) 1-NN decision boundary



(d) 10-NN decision boundary

Figure 1: Decision boundaries of the first data set

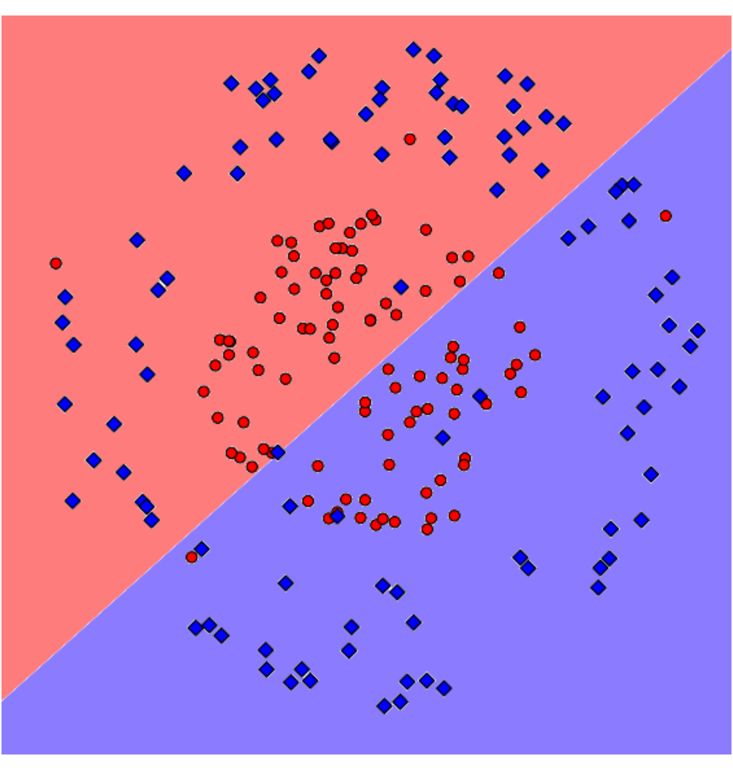
In Figure 1 we can see the decision boundaries on the first data set.

Figure 1a shows the logistic regression boundary which liear between the two sets.

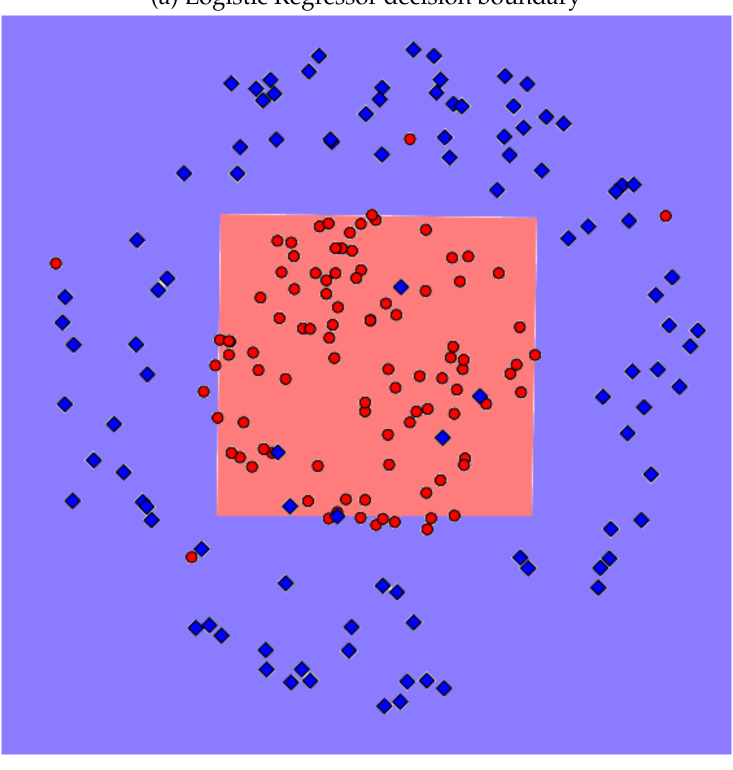
Figure 1b shows that Naive Bayes performs poorly on this data set because the features are NOT independant of each other. One feature becomes completely irrelevant to classification.

Figure 1c shows that 1-NN is very sensitive to the noise, and creates strange voronoi regions where it provides the obviously wrong label.

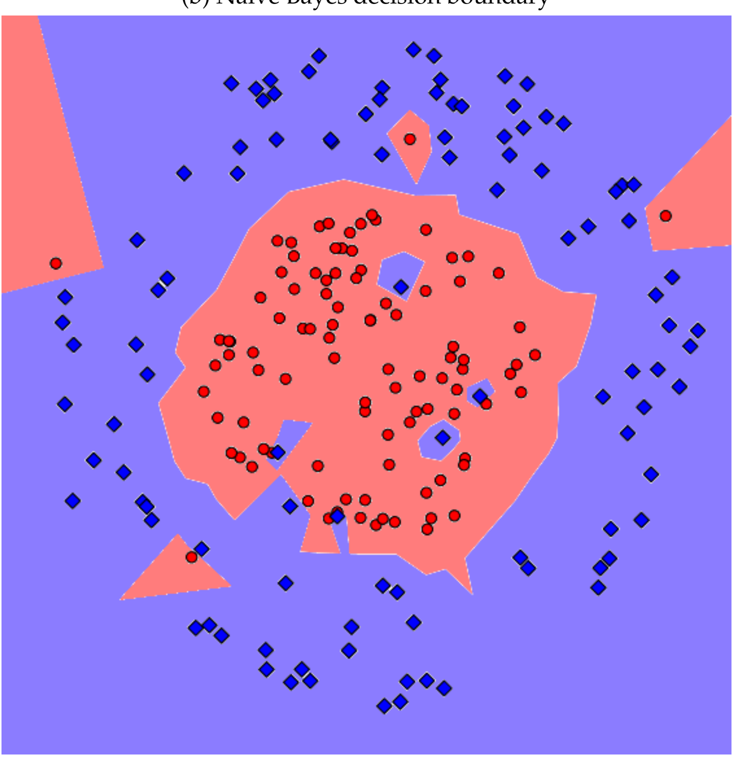
Figure 1d shows 10-NN performing very well on noise, and providing a boundary similar to the logistic regressor.



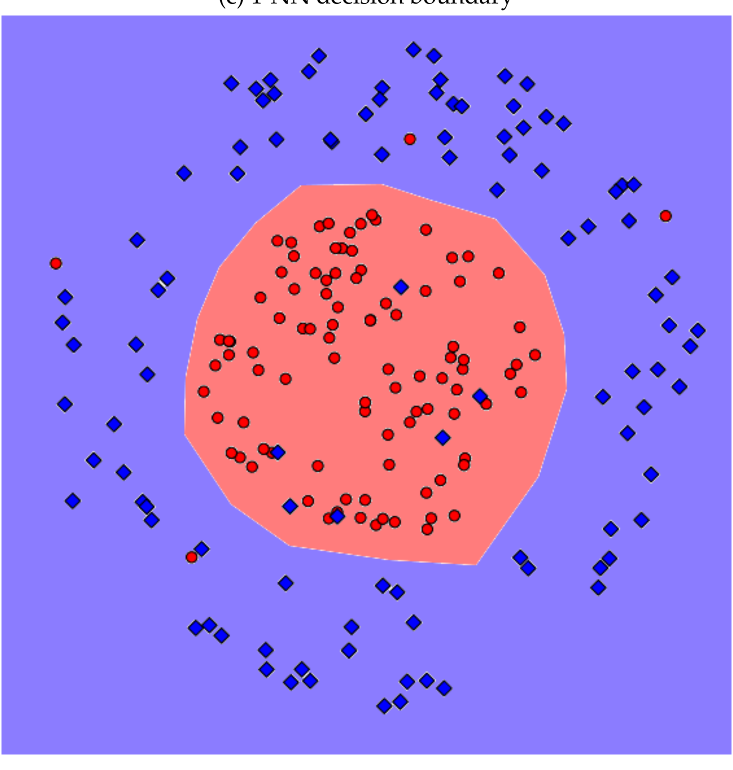
(a) Logistic Regressor decision boundary



(b) Naive Bayes decision boundary



(c) 1-NN decision boundary



(d) 10-NN decision boundary

Figure 2: Decision boundaries of the second data set

Figure 2 shows the classifier decision boundaries on the circular data set.

Figure 2a shows the logistic regression decision boundary. Note that it is unable to find a line to seperate the data. It performs poorly on this data set.

Figure 2b shows the Naive Bayes classifier decision boundary creating roughly circular decision boundary. This is because the guassian for the red-data on both features are highly kurtotic so for when posterior probability for red is high are represented by a roughly circular region.

Figure 2c shows the 1-NN classifier decision boundary. You can see that the same sensitivity to noise occurs, and huge voronoi regision of red appear on the peripherry.

Figure 2d shows the 10-NN classifier decision boundary which is robust to noise and has no patches like in 1-NN.

3.2 Defeating classifiers