Homework 2 Solution

Problem a.

Criterion	**Nearest neighbors**	**Support Vector Machine**	**Naive Bayes**
the type of classifier	non-linear	linear	linear
hypothesis space	flexible	fixed	fixed
type of learning algorithm	lazy	eager	eager
online vs. batch	online	batch	batch
robust to outliers	not robust	not robust	robust

(1) Nearest neighbors is non-linear classifier because its decision boundaries can be complex combination of segments.

(2) Support Vector Machine is linear classifier because its decision boundary is a line or hyperplane.

(3) Naive Bayes is linear classifier because Naive Bayes gives a linear decision boundary for binary feature spaces.

(4) Nearest neighbors has flexible hypothesis space because its decision boundaries can be complex combination of segments.

(5) Support Vector Machine has fixed hypothesis space because it is a linear classifier.

(6) Naive Bayes has fixed hypothesis space because it is a linear classifier in the case of multinomial formula.

(7) Nearest neighbors is lazy learning algorithm because it doesn't learn when training date come in and only predict when has testing data.

(8) Support Vector Machine is eager learning algorithm because it trains a model from training data and uses the model to predict for testing data.

(9) Naive Bayes is eager learning algorithm because it trains a model from training data and uses the model to predict for testing data.

(10) Nearest neighbors is online-learning because it can accept new data points without re-training with all the data.

(11) Support Vector Machine is batch-learning because it needs retraining with all the data every time has new data points.

(12) Naive Bayes is batch-learning because it needs retraining with all the data every time has new data points.

(13) Nearest neighbors is not robust to outliers because an outlier which is closest to a testing data point can flip its predicted class.

(14) Support Vector Machine is not robust to outliers because an outlier which is far from the decision margin can have big influence on it.

(15) Naive Bayes is robust to outliers because the count and fraction way can reduce the influence of small amount of outliers.

Problem b.

The decision boundaries are just those middle line segments between each closest pair of examples in opposite classes. After combining all the line segments, we can get the decision boundaries as shown in Figure 1.

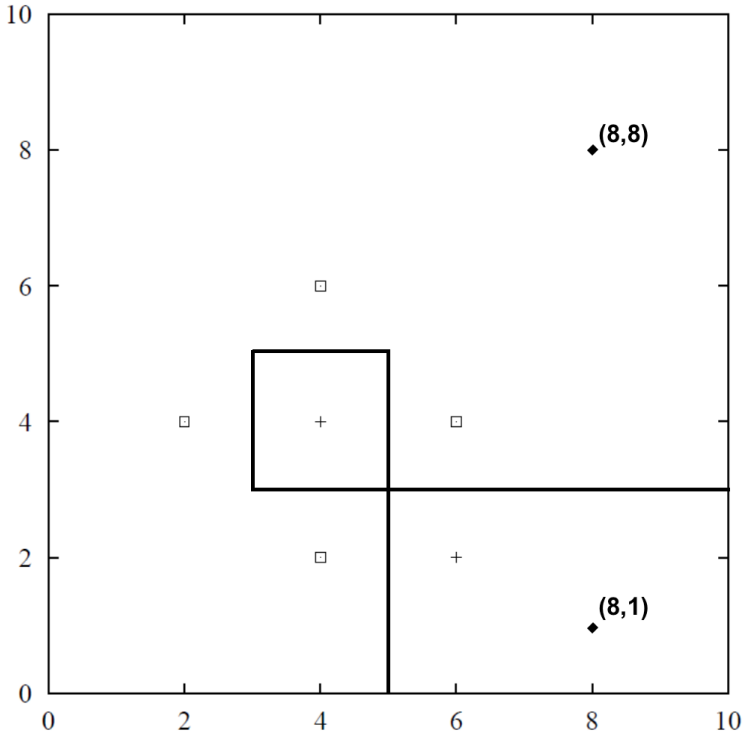

Figure 1: Decision boundaries and examples

As shown in Figure 1, we can find the location of points $(8, 1)$ and $(8, 8)$. So $(8, 1)$ is classified as positive example, and $(8, 8)$ is classified as negative example.

Problem c.

For mutual information, we know

$$I(X,Y) = H(Y) - H(Y|X)$$

So we can calculate $H(y)$,

$$\begin{aligned} H(y) &= -P(y=0)\log P(y=0) - P(y=1)\log P(y=1) \\ &= -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} \\ &= 0.9183 \end{aligned}$$

If x_1 is used to split at the root,

$$\begin{aligned} H(y|x_1) &= -P(x_1=0)[P(y=0|x_1=0)\log P(y=0|x_1=0) \\ &\quad + P(y=1|x_1=0)\log P(y=1|x_1=0)] \\ &\quad - P(x_1=1)[P(y=0|x_1=1)\log P(y=0|x_1=1) \\ &\quad + P(y=1|x_1=1)\log P(y=1|x_1=1)] \\ &= -\frac{4}{6}[\frac{2}{4}\log\frac{2}{4} + \frac{2}{4}\log\frac{2}{4}] - \frac{2}{6}[\frac{2}{2}\log\frac{2}{2} + \frac{0}{2}\log\frac{0}{2}] \\ &= 0.6667 \end{aligned}$$

The mutual information is,

$$\begin{aligned} I(x_1,y) &= H(y) - H(y|x_1) \\ &= 0.9183 - 0.6667 \\ &= 0.2516 \end{aligned}$$

If x_2 is used to split at the root,

$$\begin{aligned} H(y|x_2) &= -P(x_2=0)[P(y=0|x_2=0)\log P(y=0|x_2=0) \\ &\quad + P(y=1|x_2=0)\log P(y=1|x_2=0)] \\ &\quad - P(x_2=1)[P(y=0|x_2=1)\log P(y=0|x_2=1) \\ &\quad + P(y=1|x_2=1)\log P(y=1|x_2=1)] \\ &= -\frac{2}{6}[\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}] - \frac{4}{6}[\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}] \\ &= 0.8742 \end{aligned}$$

The mutual information is,

$$\begin{aligned} I(x_2,y) &= H(y) - H(y|x_2) \\ &= 0.9183 - 0.8742 \\ &= 0.0441 \end{aligned}$$

Since $I(x_1,y) > I(x_2,y)$, the feature x_1 will be used to split at the root.

Problem d.

For Naive Bayes Classifier,

$$P(y|x) = \frac{P(x|y)P(y)}{\sum_y P(x|y)P(y)}$$

To predict which class the example x belongs to, we only need to compare the value of $P(x|y)P(y)$ for each class and find the largest one for the predicted class.

$$\begin{aligned} P(x|y)P(y) &= P(x_1, x_2|y)P(y) \\ &= P(x_1|y)P(x_2|y)P(y) \end{aligned}$$

From the training examples, we can estimate the probabilities,

$$\begin{aligned} &P(y=0) = \frac{4}{6} = 0.6667 \qquad P(y=1) = \frac{2}{6} = 0.3333 \\ &P(x_1=0|y=0) = \frac{2}{4} = 0.5 \qquad P(x_1=1|y=0) = \frac{2}{4} = 0.5 \\ &P(x_1=0|y=1) = \frac{2}{2} = 1 \qquad P(x_1=1|y=1) = \frac{0}{2} = 0 \\ &P(x_2=0|y=0) = \frac{1}{4} = 0.25 \qquad P(x_2=1|y=0) = \frac{3}{4} = 0.75 \\ &P(x_2=0|y=1) = \frac{1}{2} = 0.5 \qquad P(x_2=1|y=1) = \frac{1}{2} = 0.5 \end{aligned}$$

So we can calculate the probabilites of features with classes,

$$\begin{aligned} P(x_1=0|y=1)P(x_2=0|y=1)P(y=1) &= 1 \times 0.5 \times 0.3333 \\ &= 0.1667 \\ P(x_1=0|y=0)P(x_2=0|y=0)P(y=0) &= 0.5 \times 0.25 \times 0.6667 \\ &= 0.0833 \\ P(x_1=1|y=1)P(x_2=0|y=1)P(y=1) &= 0 \times 0.5 \times 0.3333 \\ &= 0 \\ P(x_1=1|y=0)P(x_2=0|y=0)P(y=0) &= 0.5 \times 0.25 \times 0.6667 \\ &= 0.0833 \end{aligned}$$

Since $P(x_1=0|y=1)P(x_2=0|y=1)P(y=1) > P(x_1=0|y=0)P(x_2=0|y=0)P(y=0)$, the corresponding y of $(x_1=0, x_2=0)$ is 1.
Since $P(x_1=1|y=0)P(x_2=0|y=0)P(y=0) > P(x_1=1|y=1)P(x_2=0|y=1)P(y=1)$, the corresponding y of $(x_1=1, x_2=0)$ is 0.

Problem e.

Log likelihood objective function
Like logistic regression, since we have the log likelihood objective function $J(w)$, we will maximize it.

$$\begin{aligned} J(\mathbf{W}) &= \sum_{i=1}^{N} l(y^i|\mathbf{x}^i, \mathbf{W}) \\ l(y^i|\mathbf{x}^i, \mathbf{W}) &= \log P(y^i|A^i, \mathbf{W}) \\ &= y_1^i \log[P(y_1^i = 1|A^i, \mathbf{W})] + y_2^i \log[P(y_2^i = 1|A^i, \mathbf{W})] \\ &+ y_3^i \log[P(y_3^i = 1|A^i, \mathbf{W})] \end{aligned}$$

Compute $\frac{\partial J_i(\mathbf{W})}{\partial w_{9,6}}$.
The derivative of the given activation function $\sigma(\mathbf{x}, i) = \frac{\exp(x_i)}{\sum_{j=1}^{3} \exp(x_j)}$ is as follows. Note that $j = 1, 2, 3$ in this expression corresponds to y_1, y_2, y_3.

$$\begin{aligned} \frac{\partial \sigma(\mathbf{x}, i)}{\partial x_i} &= \frac{\exp(x_i) \cdot \sum_{j=1}^{3} \exp(x_j) - \exp(x_i)^2}{\left(\sum_{j=1}^{3} \exp(x_j)\right)^2} \\ &= \frac{\exp(x_i)}{\sum_{j=1}^{3} \exp(x_j)} - \frac{(\exp(x_i))^2}{\left(\sum_{j=1}^{3} \exp(x_j)\right)^2} \\ &= \sigma(\mathbf{x}, i) - \sigma(\mathbf{x}, i)^2 \\ &= \sigma(\mathbf{x}, i)(1 - \sigma(\mathbf{x}, i)) \\ \frac{\partial \sigma(\mathbf{x}, i)}{\partial x_k} &= \frac{-\exp(x_i) \cdot \exp(x_k)}{\left(\sum_{j=1}^{3} \exp(x_j)\right)^2} \\ &= -\sigma(\mathbf{x}, i) \cdot \sigma(\mathbf{x}, k), \text{ where } k \in \{1, 2, 3\}, k \neq i \end{aligned}$$

recall that derivative of $\frac{f(x)}{g(x)}$ is $\frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$.

$$\begin{aligned} \frac{\partial J_i(\mathbf{W})}{\partial w_{9,6}} &= y_1^i \cdot \frac{1}{P(y_1^i = 1|A^i, \mathbf{W})} \cdot \frac{\partial P(y_1^i = 1|A^i, \mathbf{W})}{\partial w_{9,6}} \ldots (1) \\ &+ y_2^i \cdot \frac{1}{P(y_2^i = 1|A^i, \mathbf{W})} \cdot \frac{\partial P(y_2^i = 1|A^i, \mathbf{W})}{\partial w_{9,6}} \ldots (2) \\ &+ y_3^i \cdot \frac{1}{P(y_3^i = 1|A^i, \mathbf{W})} \cdot \frac{\partial P(y_3^i = 1|A^i, \mathbf{W})}{\partial w_{9,6}} \ldots (3) \\ &= y_1^i \cdot (1 - \hat{y}_1^i) \cdot a_6 - y_2^i \cdot \hat{y}_1^i \cdot a_6 - y_3^i \cdot \hat{y}_1^i \cdot a_6 \\ &= [y_1^i \cdot (1 - \hat{y}_1^i) - y_2^i \cdot \hat{y}_1^i - y_3^i \cdot \hat{y}_1^i] a_6 \\ &= [y_1^i - \hat{y}_1^i (y_1^i + y_2^i + y_3^i)] a_6 \\ &= (y_1^i - \hat{y}_1^i) a_6 \end{aligned}$$

Hence,

$$\frac{\partial J(\mathbf{W})}{\partial w_{9,6}} = \sum_{i=1}^{N} (y_1^i - \hat{y}_1^i) a_6$$

Note that calculations for (1), (2), (3) are as follows.

$$\begin{aligned}
(1) &= y_1^i \cdot \frac{1}{\sigma(\mathbf{W} \cdot A^i, 9)} \cdot \frac{\partial \sigma(\mathbf{W} \cdot A^i, 9)}{\partial w_{9,6}} \\
&= y_1^i \cdot \frac{1}{\sigma(\mathbf{W} \cdot A^i, 9)} \cdot \sigma(\mathbf{W} \cdot A^i, 9)(1 - \sigma(\mathbf{W} \cdot A^i, 9)) \cdot \frac{W_9 \cdot A^i}{\partial w_{9,6}} \\
&= y_1^i \cdot (1 - \sigma(\mathbf{W} \cdot A^i, 9)) \cdot a_6 \\
&= y_1^i \cdot (1 - \hat{y}_1^i) \cdot a_6
\end{aligned}$$

$$\begin{aligned}
(2) &= y_2^i \cdot \frac{1}{\sigma(\mathbf{W} \cdot A^i, 10)} \cdot \frac{\partial \sigma(\mathbf{W} \cdot A^i, 10)}{\partial w_{9,6}} \\
&= y_2^i \cdot \frac{1}{\sigma(\mathbf{W} \cdot A^i, 10)} \cdot -\sigma(\mathbf{W} \cdot A^i, 10)\sigma(\mathbf{W} \cdot A^i, 9) \cdot \frac{W_9 \cdot A^i}{\partial w_{9,6}} \\
&= y_2^i \cdot -\sigma(\mathbf{W} \cdot A^i, 9) \cdot a_6 \\
&= -y_2^i \cdot \hat{y}_1^i \cdot a_6
\end{aligned}$$

$$\begin{aligned}
(3) &= y_3^i \cdot \frac{1}{\sigma(\mathbf{W} \cdot A^i, 11)} \cdot \frac{\partial \sigma(\mathbf{W} \cdot A^i, 11)}{\partial w_{9,6}} \\
&= y_3^i \cdot \frac{1}{\sigma(\mathbf{W} \cdot A^i, 11)} \cdot -\sigma(\mathbf{W} \cdot A^i, 11)\sigma(\mathbf{W} \cdot A^i, 9) \cdot \frac{W_9 \cdot A^i}{\partial w_{9,6}} \\
&= y_3^i \cdot -\sigma(\mathbf{W} \cdot A^i, 9) \cdot a_6 \\
&= -y_3^i \cdot \hat{y}_1^i \cdot a_6
\end{aligned}$$