CS534 — Midterm — Spring 2012

*Solutions*

## Name (Please print):

1. You have 50 minutes to finish the exam.
2. There are 6 pages in this exam (including cover page).
3. If you use the back of the page please indicate on the front of the page so I won't miss it.
4. This exam is **open book, open notes, but no cell phone and no computer**.

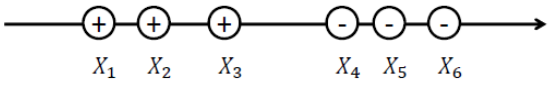
| | Max | score |
|---|---|---|
| 1 | 14 | |
| 2 | 8 | |
| 3 | 9 | |
| 4 | 19 | |
| Total | 50 | |

1. **Short Questions**.

   a. (2pts) Using a kernel function is equivalent to mapping data into a higher dimensional space then taking the linear dot product in that space. Please explain what is the advantage of using the kernel function compared to doing the explicit mapping. *Solution: The main advantage is computational complexity. It also allows us to map to an infinite dimensional space when using RBF kernel, which is not possible via explicit mapping.*

   b. (2pts) Prove that $P(A|B)P(B) = P(B|A)P(A)$.
   *Solution:* $P(A|B)P(B) = P(A, B) = P(B|A)P(A)$

   c. (4 pts) Consider applying a soft margin SVM to the 1-dimensional data shown below.




   What will be the support vectors for $c = 0$ and $c = \infty$ respectively?

   $C = 0$: *All points will be support vectors. In this case, there is no penalty for positive $\xi$'s, thus the learned boundary will place all points inside the two lines of $\mathbf{w}^T\mathbf{x} + b = 1$ and $\mathbf{w}^T\mathbf{x} + b = -1$, making every point a support vector.*

   $C = \infty$: *$X_3$ and $X_4$. This is equivalent to hard margin SVM because any positive $\xi$ will cause infinitely large penalty, thus strictly avoided.*

   d. (2pts) **True or false. Provide a brief explanation.** Consider applying Naive Bayes to a classification problem. Suppose the modeling assumptions made by Naive bayes are true, and we have infinite training data, the learned Naive Bayes classifier will have zero training error.
   *Flase. When modeling assumption is correct, a generative model like Naive Bayes can learn optimal decision boundary with infinite training data. However, it may not achieve zero training error if the two classes overlap.*

e. (4pts) What impact will the following operation have on overfitting, increase, decrease or no impact?

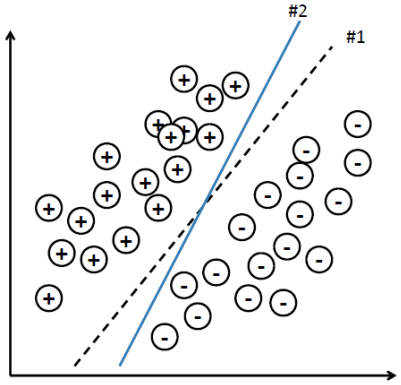
Increase $k$ for the $k$-nearest neighbor classifier: ***Decrease***

Increase $c$ for soft margin support vector machines: ***Increase***

Increase the amount of training data for Logistic regression: ***Decrease***

Remove non-support vector instances in the training set for SVM: ***no impact***

2 (2 pts each) **Linear decision boundaries**. Consider the following binary classification problem as shown in the figure below. In the figure, we provide two possible linear decision boundaries. Please specify for each of the following four algorithms, will it produce boundary #1, #2 or possibly both? Please provide a one-sentence explanation to your answer.




**Linear discriminant analysis:** #1 - *The two classes appear to be Gaussian with the same covariance structure. Under such conditions, LDA learns the boundary that separates the two distributions optimally.*

**Logistic regression:** *Both are possible depending on the initial weight and the learning rate of gradient ascent.*

**Support vector machine:** #1, *as it achieves the maximum margin.*

**Perceptron (stochastic gradient descent):** *Both are possible depending on the initial weight and the learning rate of gradient descent, and the order it receives training data.*

3 **Linear regression.** Consider linear regression using polynomial basis functions of order $M$. The expected loss can be decomposed into three parts, the bias, the variance and the noise.

a. (3pts) Please provide the expression for these three components.
*See linear regression slide* 36.

b. (3 pts) If we increase the order of the polynomial basis function (e.g., from quadratic to cubit, or higher order polynomials), will it increase, decrease or have no impact on each of the three components? Briefly explain.
*Noise: no impact because it is inherent to the data, and has nothing to do with the learning algorithm.*
*Bias: decrease because the increased model flexibility with higher order $M$ allows better fit of the data.*
*Variance: increase because the increased flexibility makes it easier to fit the particularity of the training data $D$, leading to larger variance.*

c. (3 pts) If we increase the training set size, will it increase, decrease or have no impact on each of the three components? Briefly explain.
*Noise: no impact because the term $D$ does not appear in the noise expression at all.*
*Bias: we don't expect bias to change. As we increase the training data set size, the output model does not change much* ***in expectation****. In particular, if $D$ grows to infinitely large, the output model will converge to its expectation. The bias will remain the same.*
*Variance: decrease because larger training set size makes it less likely to overfit to the particularity of $D$, as we grow $D$ to infinitely large, the variance will decrease to zero.*

4 **Naive Bayes.** Consider a binary classification problem with variable $X_1 \in \{0, 1\}$ and label $Y \in \{0, 1\}$. The true generative distribution $P(X_1, Y) = P(Y)P(X_1|Y)$ is shown in Table 1 and Table 2.

| $Y = 0$ | $Y = 1$ |
|---|---|
| 0.8 | 0.2 |

Table 1: $P(Y)$

| | $X_1 = 0$ | $X_1 = 1$ |
|---|---|---|
| $Y = 0$ | 0.7 | 0.3 |
| $Y = 1$ | 0.3 | 0.7 |

Table 2: $P(X_1|Y)$

a. (4 pts): Now suppose we have trained a Naive Bayes classifier, using infinite training data generated according to Table 1 and Table 2. Please fill in Table 3. In particular, please fill in the probabilities in the first two columns, and fill in the prediction of $Y$ in the last column of the table. For the probabilities, please write down the actual values (and the calculation process if you prefer, e.g., $0.8 \times 0.7 = 0.56$).

| | $\hat{P}(X_1, Y = 0)$ | $\hat{P}(X_1, Y = 1)$ | $\hat{Y}(X_1)$ |
|---|---|---|---|
| $X_1 = 0$ | $0.8 \times 0.7 = 0.56$ | $0.2 \times 0.3 = 0.06$ | 0 |
| $X_1 = 1$ | $0.8 \times 0.3 = 0.24$ | $0.2 \times 0.7 = 0.14$ | 0 |

Table 3: Predictions of the trained Naive Bayes

b (3 pts) What is the expected error rate of this classifier on training examples generated according to Table 1 and Table 2? In other words, what is $P(Y \neq \hat{Y}(X_1))$?
(Hint: $P(Y \neq \hat{Y}(X_1)) = P(Y \neq \hat{Y}(X_1), X_1 = 1) + P(Y \neq \hat{Y}(X_1), X_1 = 0))$
$P(Y \neq \hat{Y}(X_1) = P(Y = 1, X_1 = 0) + P(Y = 1, X_1 = 1) = 0.06 + 0.14 = 0.2$

c. Now we add a feature to this data $X_2$ such that $X_2$ is an exact duplicate of $X_1$. Suppose we have trained Naive Bayes classifier using infinite training data that are generated by following Tables 1-2, and then add the additional duplicate feature $X_2$. Please fill in the following tables.

i. (2pts) Fill in the probabilities for $P(X_2|Y)$ in Table 4.

ii. (4 pts) Fill in the probabilities for Table 5 and write down the predictions of $Y$ for different $X_1$ and $X_2$ value combinations.

| | $X_2 = 0$ | $X_2 = 1$ |
|---|---|---|
| $Y = 0$ | *0.7* | *0.3* |
| $Y = 1$ | *0.3* | *0.7* |

Table 4: Probability estimation for $P(X_2|Y)$.

| | $\hat{P}(X_1, X_2, Y = 0)$ | $\hat{P}(X_1, X_2, Y = 1)$ | $\hat{Y}(X_1, X_2)$ |
|---|---|---|---|
| $X_1 = 0, X_2 = 0$ | $0.8 \times 0.7 \times 0.7 = 0.392$ | $0.2 \times 0.3 \times 0.3 = 0.018$ | *0* |
| $X_1 = 1, X_2 = 0$ | $0.8 \times 0.7 \times 0.3 = 0.168$ | $0.2 \times 0.7 \times 0.3 = 0.042$ | *0* |
| $X_1 = 0, X_2 = 1$ | $0.8 \times 0.7 \times 0.3 = 0.168$ | $0.2 \times 0.3 \times 0.7 = 0.042$ | *0* |
| $X_1 = 1, X_2 = 1$ | $0.8 \times 0.3 \times 0.3 = 0.072$ | $0.2 \times 0.7 \times 0.7 = 0.098$ | *1* |

Table 5: Predictions of the trained Naive Bayes.

d. (3 pts) What is the expected error rate of this Naive Bayes classifier on this data?
$P(Y \neq \hat{Y}(X_1, X_2))$
$= P(Y \neq \hat{Y}(X_1, X_2), X_1 = 0) + P(Y \neq \hat{Y}(X_1, X_2), X_1 = 1)$
$= P(Y = 1, X_1 = 0) + P(Y = 0, X_1 = 1)$
$= 0.06 + 0.24 = 0.3$

e. (3 pts) Compare the error rate in *d* to the error rate in *b*. What is the reason for the difference?
*The error rate is increased. This is because $X_1$ and $X_2$ are not conditional independent given $Y$, violating the the Naive Bayes assumption.*