CS6360: Advanced Topics in Machine Learning
Total Marks: 30

January 5, 2016

Roll No: Name:

CGPA:

If any earlier ML course taken, grade and semester:

If any earlier ML project done, details:

1. *(1 × 6 = 6 marks)* State whether TRUE or FALSE (**Every wrong answer carries negative 0.5 marks**):
   (a) Decision trees are generative classifiers. T (F)
   (b) Since classification is a special case of regression, logistic regression is a special case of linear regression. T (F)
   (c) We can get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent. T (F)
   (d) When the hypothesis space is richer, overfitting is more likely. (T) F

(e) When the data is not completely linear separable, the linear SVM *without slack variables, i.e. hard margin* returns $\mathbf{w} = \mathbf{0}$. T Ⓕ

(f) After training a SVM, we can discard all examples which are not support vectors and can still classify new examples. Ⓣ F

2. **k-Nearest Neighbors (6 marks):**

(a) (2 marks) Which of the following statements are true for k-NN classifiers (choose all answers that are correct).

i. The classification accuracy is better with larger values of k.
ii. The decision boundary is smoother with smaller values of k.
iii. k-NN is a type of instance-based learning.
iv. k-NN does not require an explicit training step.

SOLUTION: (iii) and (iv)

(b) (2 marks) Is it possible for a 2-class 1-NN classifier to always classify all new examples as positive even though there are negative examples in the training data? If yes, show an example. If no, briefly explain.

SOLUTION: No. Pick a negative example and its closest positive example. Any example on the line connecting them but closer to the negative example must be classified as negative.

(c) (2 marks) Suppose we have a large training set. Name a drawback when using a k Nearest Neighbor during testing.

SOLUTION: k-NN is slow in testing phase, since the time complexity for finding k nearest neighbors is $O(knd)$. n is number of training data points. d is number of dimensions.

3. **Naive Bayes Classifier (6 marks):**

(a) (2 marks) Suppose that in answering a question from a multiple choice test, an examinee knows the answer with probability p, or he guesses with probability $1 - p$. Assume that the probability of answering a question is 1 for an examinee who knows the answer and $\frac{1}{m}$ for the examinee who guesses, where m is the number of multiple choice alternatives. What is the probability that an examinee knew the answer to a question, given that he has correctly

answered it?
SOLUTION:

$$P(\text{Know answer} \mid \text{correct}) = \frac{P(\text{know answer, correct})}{P(\text{correct})}$$

$$= \frac{p}{p + (1-p)\frac{1}{m}}$$

(b) (2+2 marks) Annabelle Antique is a collector of old paintings. She is sick of getting e-mails offering her fake artwork and wants to train her own Naive Bayes classifier such that she doesnt have to read all the spam any longer. One of the words she knows that corresponds to a fake is the occurrence of the word *replica*. The Naive Bayes classifier doesnt know this yet. Your job is to help it by generating suitable messages:

i. Generate two messages corresponding to non-spam and spam respectively, which will lead to the classifier correctly recognizing *replica* as spam. Explain your answer.

SOLUTION:
Non-spam 1: Van Gogh's Starry night in canvas
Non-spam 2: The original copy of Da Vinci's Vitruvian Man
Spam 1: Van Gogh's Starry night oil painting replica
Spam 2: Da Vinci's Vitruvian Man adapted to oil painting

ii. Generate a message that would be incorrectly classified as non-spam based on the four messages generated in your earlier answer. Explain why.

SOLUTION:
An exquisite copy of Van Gogh's Starry night

4. **Linear Regression (6 marks):**

(a) (2 marks) We would like to use the following regression function:

$$y = w^2 x + wx$$

where x is a single-valued variable. Given a set of training data points $\{(x_i, y_i)\}$, derive a solution for w (an expression for the solution will do).

SOLUTION:

$$(w^2 + w) = \sum_i \frac{x_i y_i}{x_i^2}$$

Note that the model is equivalent to

$$y = (w^2 + w)x$$

So, the solution for $(w^2 + w)$ will be exactly the same as a simple linear $y = wx$ model.

(b) (2+2 = 4 marks) We would like to compare the regression model used in the earlier sub-question to the following regression model:

$$y = wx$$

i. Given limited training data, which model would fit the *training data* better? Explain your answer briefly.

A. Model 1

B. Model 2

C. Both will fit the data equally well

D. Impossible to tell

SOLUTION:
3. As mentioned above, the solution to the optimization problem is the same and so the outcome will be the same as well (any value that can be expressed using w can be expressed using (w 2 + w) and so the two models are exactly the same.

ii. Given limited training data, which model would fit the *test data* better? Explain your answer briefly.

A. Model 1

B. Model 2

C. Both will fit the data equally well

D. Impossible to tell

SOLUTION:
3. Same reason as above.

5. **Support Vector Machines (6 marks):**

(a) (2 marks) Prove that the kernel $K(x_1, x_2)$ is symmetric, where x_i and x_j are the feature vectors for i^{th} and j^{th} examples. (Hint: Your answer should not be more than 2-3 lines).

SOLUTION: Let $\phi(x_1)$ and $\phi(x_2)$ be the feature maps for x_i and x_j , respectively. Then, we have $K(x_1, x_2) = \phi(x_1)'\phi(x_2) = \phi(x_2)'\phi(x_1) = K(x_2, x_1)$

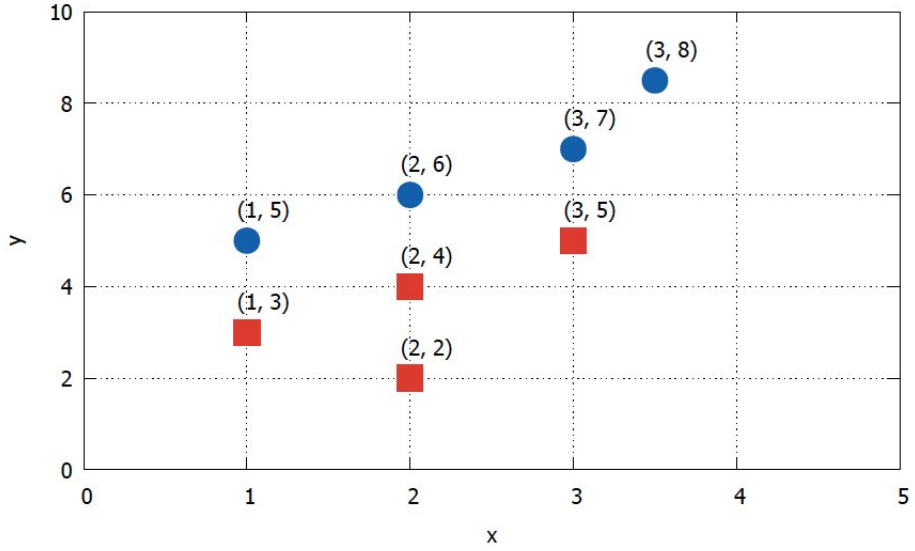
(b) (2 marks) Consider soft margin SVM, described as the following optimization problem:

$$arg \min_{W,b} |W|^2 + C \sum_i \Xi_i$$

subject to $y_i.(W.X_i + b) + \Xi_i \geq 1$ and $\Xi_i \geq 0$ for $i = 1, \cdots, n$. This can be viewed as a penalty-based method to avoid overfitting. Please specify which term in the objective function is the penalty term and which is the error term.

SOLUTION: Ξ = Error; C = penalty

(c) (2 marks) The following figure shows few sample points for a classification problem. The points marked with circles are from one class, and the points marked with squares are from another class. What is the equation of the Support Vector Classifier (hard margin, no kernel) for the data?


10
8
6
4
2
0
y
0
1
2
3
4
5
x
(3, 8)
(3, 7)
(2, 6)
(1, 5)
(3, 5)
(2, 4)
(1, 3)
(2, 2)