CS534 — Midterm — Fall 2018

12pm - 13:20pm Thursday, Nov 1st

Name (Please print): Solution

1. You have 80 minutes to finish the exam.
2. There are 7 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 we won't miss it.
4. This exam is **close book, close notes, but you are allowed one page cheat sheet (letter size, one sided)**. At the end of the exam, submit the cheat sheet with your exam paper.
5. Electronic devices (including calculator) are not allowed.

	Max	score
1	23	
2	10	
3	10	
4	6	
5	20	
Total	69	

1. (23 pts) Short questions.

   (i) (3pts) Given a weighted coin with unknown probability p for head. We toss the coin N times and observe n_1 heads and n_0 tails. What is the likelihood function?

   $$L(p) = p^{n_1}.(1-p)^{n_0} \tag{1}$$

   (ii) (4 pts) Let x_1 and x_2 denote the outcome of two coin tosses respectively. Consider the following statements, check all statements that are true.

   ✓ x_1 and x_2 are independent of each other.

   ✓ $p(x_1|x_2) = p(x_1)$

   ✗ $p(x_2|x_1) = p(x_1)$

   ✗ x_1 and x_2 are conditionally independent regardless of the condition.

   (iii) (2 pts) Given a multiple choice question, students either know the answer, with probability p, or they don't know and guess, with probability $1-p$. Further if a student knows the answer, the probability of correct is 1, otherwise, the probability of correct is $1/m$, where m is the number of choices. What is the probability that a student knows the answer given that she/he has correctly answered the question?

   a. [✓] $\frac{p}{p+(1-p)\times\frac{1}{m}}$

   b. p

   c. $p + (1-p) \times \frac{1}{m}$

   d. $p \times (1 - \frac{1}{m})$

   (iv) (4 pts) Consider linear regression with the sum of squared error (SSE) objective $\sum_{i=1}^{N}(w^T x_i - y_i)^2$. Which of the following statements are true? Select all that are correct.

   ✓ We assume that all examples in the training set are IID - identically and independently distributed.

   ☐ We assume that the target y is a linear function of the input x plus a Gaussian random noise, where the variance of the noise can be different for all examples.

   ✓ The Sum of Squared Error objective (SSE) is equivalent to the mean squared error objective (MSE) $\frac{1}{N}\sum_{i=1}^{N}(w^T x_i - y_i)^2$

   ☐ Because the linear function is super simple, there is no risk of over-fitting when using the linear regression model.

   (v) (2pts) Consider a binary classification problem with d binary features. What is the total number of parameters that the Naive Bayes Classifier need to learn?

   a. $(2^d - 1) \times 2 + 1$

   b. $2 \times (d-1) + 1$

c. $2 \times d$

✓ [d.] $2 \times d + 1$

(vi) (4 pts) Consider the convergence theorem for the perceptron algorithm. Which of the following statements are true? Select all that are correct.

✓ The perceptron algorithm will always converge in a finite number of steps if the training data is linearly separable.

☐ When the data is linearly separable, the perceptron algorithm will always converge to linear separator that has the largest margin.

☐ The convergence rate is independent of the order of the examples during training

☐ Because the number of updates k is bounded by a quantity proportional to D, the largest norm of input x, we can reduce the number of udpates needed to converge by normalizing the data to have norm of 1.

(vii) (4 pts) Which of the following statements are true about overfitting? Select all that are true.

☐ The more training data we have, the more likely we will overfit.

✓ The more features we have, the more likely we will overfit.

✓ When using SVM with RBF kernel ($K(x_1, x_2) = \exp(-\frac{|x_1 - x_2|^2}{2\sigma})$), the smaller the σ, the more likely we will overfit.

✓ When using L_2 regularization, the larger the regularization parameter λ the less likely we will over fit.

2. (10pts) Naive Bayes Consider the following training examples (x, y).

$x^1 = (0,0,0,1,0,0,1) \qquad y^1 = 1$
$x^2 = (0,0,1,1,0,0,0) \qquad y^2 = 1$
$x^3 = (1,1,0,0,0,1,0) \qquad y^3 = -1$
$x^4 = (1,0,0,0,1,1,0) \qquad y^4 = -1$

(a) (3 pts) What problem will Naive Bayes encounter with test data $x = (1,0,0,0,0,0,0)$? Probabilities $p(x|y=1)$ and $p(x|y=-1)$ are both 0. Therefore $p(y|x)$ is zero for both.

(b) (7 pts) Now apply Laplace smoothing when estimating $P(x_i|y)$ for $i = 1, ..., 7$, where x_i is the i-th feature of example x, and compute $P(y=1|x)$ using Naive Bayes.

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1) + p(x|y=-1)p(y=-1)} \tag{2}$$

$$= \frac{(\frac{1}{4}\frac{3}{4}\frac{1}{2}\frac{1}{4}\frac{3}{4}\frac{3}{4}\frac{1}{2})\frac{1}{2}}{(\frac{1}{4}\frac{3}{4}\frac{1}{2}\frac{1}{4}\frac{3}{4}\frac{3}{4}\frac{1}{2})\frac{1}{2} + (\frac{3}{4}\frac{1}{2}\frac{3}{4}\frac{3}{4}\frac{1}{2}\frac{1}{4}\frac{3}{4})\frac{1}{2}} \tag{3}$$

$$= \frac{1}{4} \tag{4}$$

3. (10 pts) Support Vector Machine. Consider the following figure, which plots the decision boundaries resulting from SVM using different kernels with different parameters. The data contains two classes, denoted by squares and circles respectively. The solid squares and circles represent the support vectors. Please label each image with the corresponding kernel and parameter choice (among the provided options i-v) that produces the decision boundary and support vectors as depicted in the figure.

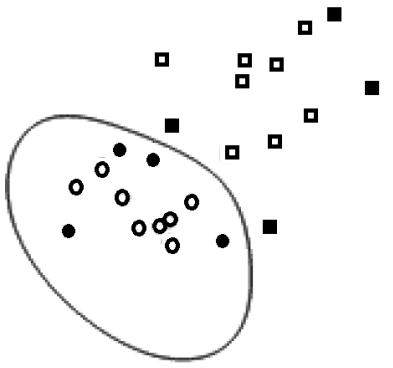
   (i) A soft margin SVM with linear kernel and $c = 0.01$ (e)

   (ii) A soft margin SVM with linear kernel and $c = 100$ (b)

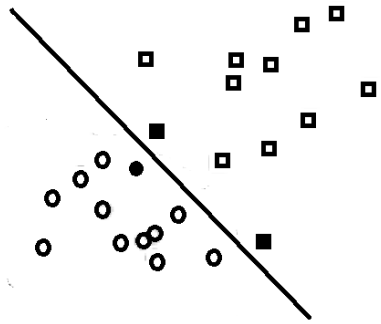
   (iii) A hard margin SVM with a quadratic kernel (c)

   (iv) A hard margin SVM with RBF kernel $K(x_1, x_2) = \exp(-\frac{|x_1 - x_2|^2}{2\sigma})$ with $\sigma = 2$ (a)

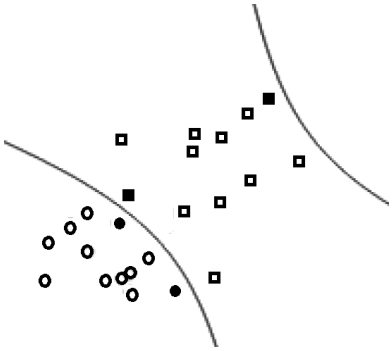
   (v) A hard margin SVM with RBF kernel $K(x_1, x_2) = \exp(-\frac{|x_1 - x_2|^2}{2\sigma})$ with $\sigma = 0.5$ (d)



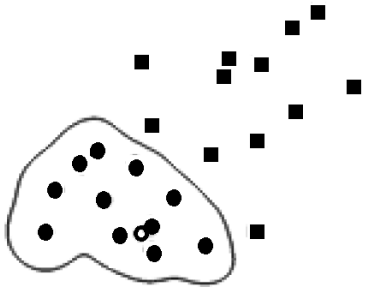
(a)



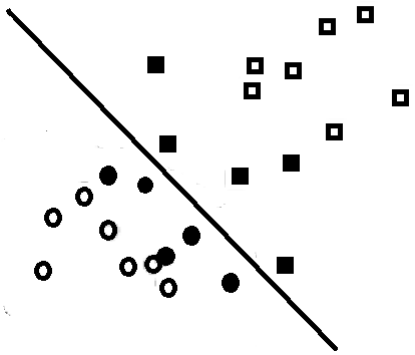
(b)



(c)

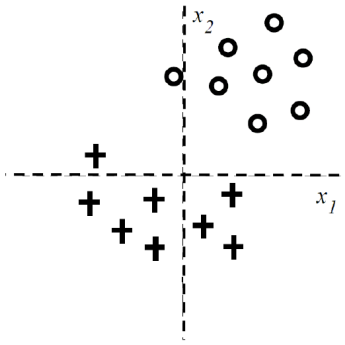


(d)



(e)

4. (6pts) Logistic regression Consider the following binary classification problem.



We will train a regularized logistic regression model $P(y = 1|\mathbf{x} = (x_1, x_2)) = \frac{1}{1+\exp(-w_0-w_1x_1-w_2x_2)}$ by maximizing the following objective:

$$\sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i; w_0, w_1, w_2) + \lambda w_j^2$$

where the regularization term only penalize one of the coefficients w_j, $j \in \{0, 1, 2\}$. With different choice of j, we expect different behavior from the learner. Assume that we are using a very large λ, please state whether the training error will increase for each of the following cases, with a brief justification.

(a) (2pts) Regularizing w_0.
Error will increase. With a heavy regularization of w_0, the model learns a weight w representing a line passing through the origin. No such a line can separate entire samples.

(b) (2pts) Regularizing w_1.
Error will not increase. With a heavy regularization of w_1, the model learns a weight w representing a horizontal line. There exists a horizontal line to separate entire samples.

(c) (2pts) Regularizing w_2.
Error will increase. With a heavy regularization of w_2, the model learns a weight w representing a vertical line. There is no vertical line to separate the entire samples.

5. (20pts) True/False. No explanation needed.

(i) (True/False) When using gradient descent to train a linear regression model to minimize the Sum of Squared Error with L_2 regularization, different initializations will lead to the same global minimum.

(ii) (True/False) The voted perceptron algorithm learns a linear decision boundary.

(iii) (True/False) The average perceptron algorithm learns a linear decision boundary.

(iv) (True/False) If K_1 and K_2 are two kernel functions, then $\alpha_1 K_1 + \alpha_2 K_2$ will also be a kernel for any α_1 and α_2.

(v) (True/False) Removing non-support vectors from the training set will not impact the learned decision boundary for SVM.

(vi) (True/False) A two layer neural network can represent the XOR function.

(vii) (True/False) Feature engineering is not important when we use kernels because they map the data to a nonlinear space in which data is linearly separable.

(viii) (True/False) For Maximum A Posterior estimation, if we have infinite amount of data, the influence of the prior will always be neglectable regardless of the prior.

(ix) (True/False) The minimum number of support vectors for a two-class dataset is 2.

(x) (True/False) If the training data is linearly separable, the classifier learned by linear SVM will not misclassify any training examples.