CS5489 Quiz
Semester B 2020

Instructions:

- Answer all questions in this Word document.
- After each question, put your answer in the space marked "ANSWER".
- The following resources are **allowed** during the quiz:
  - Panopto lecture videos (or the same videos on OneDrive),
  - any material on the CS5489 Canvas page, including lecture notes, tutorials,
  - course textbooks
  - videos on Zoom taken during the lecture discussions.
- Any other resources are **not allowed**, for example
  - internet searches
  - classmates
  - other textbooks
- You should stay on Zoom during the entire quiz time in case there are any announcements.
  - If you have any questions, please use the private chat function of Zoom to message Antoni.
- By 9pm March 2nd, submit the completed quiz to the Midterm Assignment on Canvas.
  - If you have trouble accessing Canvas, then you can send the completed quiz through private Zoom Chat or through Email to Antoni.

Statement of Academic Honesty

I pledge that the answers in this exam/quiz are my own and that I will not seek or obtain an unfair advantage in producing these answers. Specifically,

- *I will not plagiarize (copy without citation) from any source;*
- *I will not communicate or attempt to communicate with any other person during the exam/quiz; neither will I give or attempt to give assistance to another student taking the exam/quiz; and*
- *I will use only approved resources during the quiz. I understand that any act of academic dishonesty can lead to disciplinary action.*

By putting my name below, I reaffirm my academic honesty pledge.

Name: ANSWER KEY
EID: <PUT YOUR EID HERE>
Student ID: <PUT YOUR STUDENT ID HERE>

Multiple Choice/Selection Questions (30 points)
5 marks each. For multiple selection questions, all answers need to be correct.

Q1. Which statements are not true about Logistic Regression? (select all that apply)
A) When training a Logistic Regression model, we have only one local optimum solution.
B) When training a Logistic Regression model, we should fix the learning rate (η) to obtain the local optimum.
C) When training a Logistic Regression model with L2-norm regularization, while increasing the hyperparameter C, the training error will increase and testing error will decrease.
D) When performing classification with Logistic Regression, we can use different decision function (e.g., sigmoid, sign) in training and testing process.

Q1 ANSWER: <B, C>

Q2. Suppose you have trained a classifier for binary classification task. It gives high accuracy on the training set, but the accuracy is low on the test set. Which methods could be adopted to improve the test accuracy? (select all that apply)
A) Delete some training samples randomly.
B) Employ more complex classifiers.
C) Increase the regularization.
D) Employ Cross-validation on the training set.
E) Train an ensemble of classifiers using bagging.

Q2 ANSWER: <C, D, E>

Q3. Which statements about Gaussian Process regression (GPR) are correct? (select all that apply)
A) GPR is defined as the Bayesian linear regression whose linear kernel is replaced by the Gaussian/RBF kernel.
B) GPR only has a closed form solution when the observation noise is Gaussian.
C) One assumption in the Gaussian process prior is that two function values are more correlated when the corresponding inputs are close together.
D) The reason why GPR is not good for large datasets is that it only has a few parameters and thus limited model complexity.
E) GPR assumes the input data points are i.i.d. (independently and identically distributed).

Q3 ANSWER: <B, C>

Q4. Which statements about cross-validation are correct? (select all that apply)
A) The validation set for cross-validation is a part of the training data.

B) The best parameters are found through comparing the performance on the testing set.
C) In logistic regression, the corresponding w and b of the best C in the cross-validation stage might be different from the w and b in the training stage with the best C.
D) If the final testing score is still low with the best parameters selected through cross-validation, the reason could be that the step for the parameter selection is too large.

Q4. ANSWER: <A, C, D>

Q5. Which statements about the naive Bayes classifiers are correct? (select all that apply)
A) Learning a Bayesian classifier is equivalent to estimating the posterior distribution.
B) A Bayes classifier that has a large prediction variance around the decision boundary has a good uncertainty property.
C) If the generative distribution is not Gaussian, the naive Bayesian classifier can be nonlinear.
D) Both naive Bayes classifier and logistic regression can learn model parameters by maximum likelihood estimation.
E) There is no regularization in a naive Bayes classifier.

Q5. ANSWER: <A, B, C, D>

Q6. Which statements are true about bagging and boosting? (select all that apply)
A) Bagging and boosting are both ensemble methods.
B) Bagging is based on iteratively building classifiers, while boosting makes the classifiers faster.
C) Bagging and boosting are only based on decision trees or decision stumps.
D) Bagging and boosting both overfit when adding too many classifiers.
E) Bagging and boosting can be combined together.

Q6. ANSWER: <A, E>

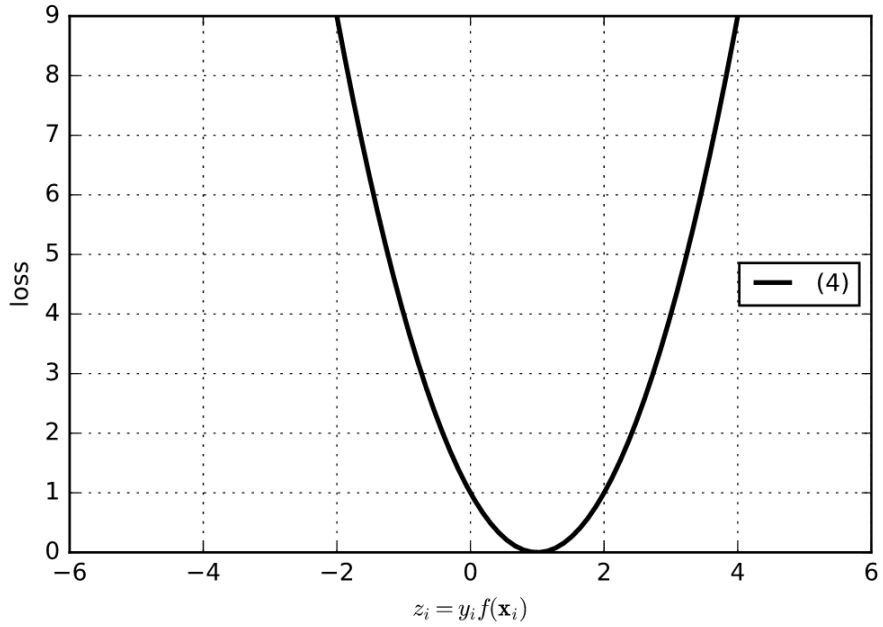
Discussion Questions (70 marks)
10 marks each question.

Q7. You have been asked to build a classifier for an early screening test for diagnosing cancer. After the screening test, a more expensive advanced test will be used to confirm the result. For the screening test, you are given a feature vector, and the goal is to predict whether or not the person has cancer (positive or negative). Since it is a screening test, the classifier should classify all people with cancer as positive, while it is okay for some people without cancer to be misclassified as positive. Discuss two methods for getting this desired behavior from the classifier.

Q7 ANSWER:
<ANSWER HERE>

- 1) use weights on the classes during training. The points in the positive class should have higher weight than the negative class, so that the classifier focuses more on predicting the positive class correctly.
- 2) change the threshold of the classifier. Usually it is T=0, but we can set it to T<0 to make it predict more positive examples.

Q8. Consider a classifier with the loss function in the below figure. Using this loss function, describe some expected properties of this classifier and explain why. Will this be a good classifier or a bad classifier? Why?



Q8 ANSWER:
<ANSWER HERE>

- The minimum at z=1 will try to make the classifier predict f(x)=+1 or f(x)=-1 only.
- Increasing loss when z<1, means it will be not robust to outliers.
- Increasing loss when z>1, means it will also try to push well-classified examples back to f(x)=1. Sensitive to "easy" or "too correct" examples.
- This will not be a good classifier.

Q9. Your friend is training a linear SVM on a patient data from a hospital. There are 2000 patients, and each patient has 100,000 features extracted from their time in the hospital. Your friend complains that training the linear SVM takes too long. What suggestions do you give your friend to speed up the training?

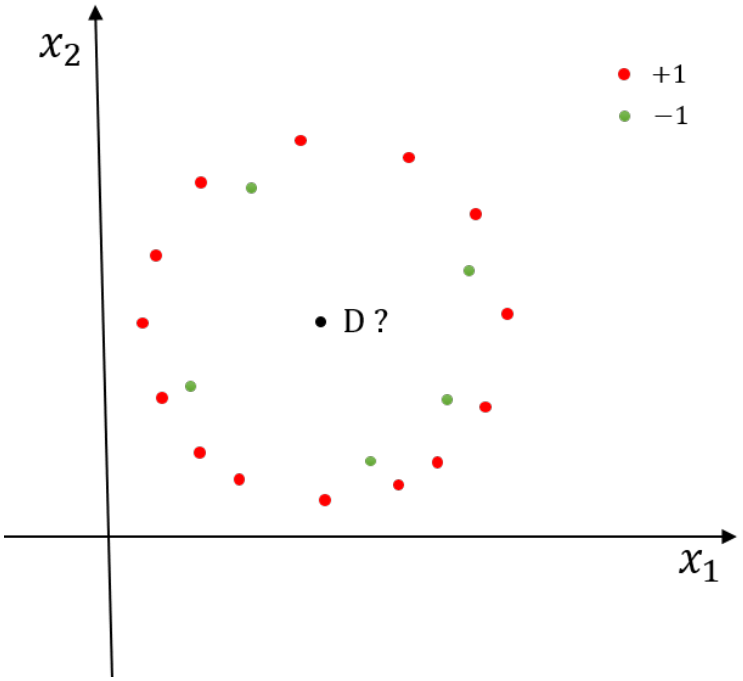
Q9 ANSWER:
<ANSWER HERE>

- If they are learning the linear SVM with the primal problem, then there are too many variables to optimize (w is 100,000 dimensions). Convert the problem into the dual

form, and they only need compute the kernel matrix once, which is 2000x2000, and the optimization is over only 2000 variables.

- Use liblinear instead of libsvm. (partial credit, if just this answer without any explanation)
- Use dimensionality reduction (e.g. PCA) – also got credit for this, although it was not introduced yet.

Q10. Consider the data distribution shown in the figure below. If you train a Bayes Gaussian classifier, which class will be predicted for data point D? If we change the classifier to an SVM with polynomial kernel of degree 2, which class will be predicted? Explain why.



Q10 ANSWER:

<ANSWER HERE>

- For Bayes Classifier, the two classes have the same centroid, while the class {+1} has larger variance in each dimension and its prior is much larger than class {-1}, so the prediction of D is {+1}.
- For SVM with poly degree-2 kernel, the prediction is {-1}, since the SVM finds the margin between the +1 and -1 class.

Q11. What is overfitting? How do we know when it happens? Write down and describe several methods to address the problem of overfitting.

Q11 ANSWER:

<ANSWER HERE>

- Overfitting happens when there are too many model parameters for the amount of training data, and the learned function becomes too complex and follows the data to well.
- The error on the training set is much lower than the error on the test/validation set.

- Overfitting can be addressed by:
  - using cross-validation to select the model hyperparameters
  - using a suitable amount of regularization
  - reducing the number of features manually
  - reducing the complexity of the class of functions being learned (e.g., changing degree-5 polynomial to degree-2)

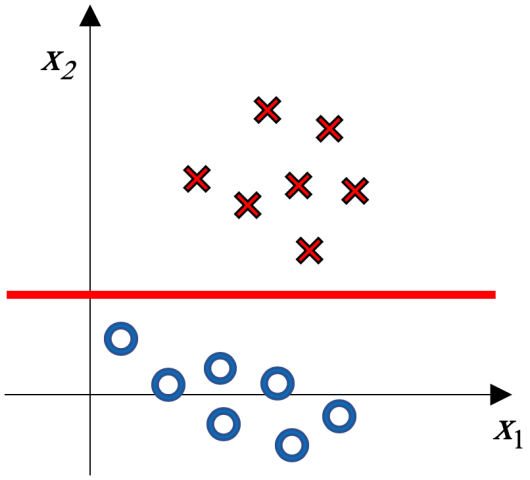
Q12a. Consider a logistic regression classifier using L1 regularization instead of L2 regularization? Write down the optimization problem for learning the classifier.

Q12a ANSWER:
<ANSWER HERE>

- $w^*, b^* = argmin_{w,b} \frac{1}{C} ||w||_1 + \sum_{\{i=1\}}^{N} log\ (1 + exp(-y_i f(x_i)))$

Q12b. Suppose we have the data in the figure below. Draw the most likely decision boundary for the L1-regularized logistic regression classifier. You can move/reshape the red line provided below. Explain why this should be the decision boundary.

Q12b ANSWER:
<ANSWER HERE>
The L1 norm on the weights will encourage w1=0 or w2=0. This means the decision boundary is aligned with the coordinate axes. I.e., the L1 norm will select the important feature for the classifier, and ignore other features.



Q13. Your friend has designed a new regression algorithm that minimizes the L1 norm of the error and L2 norm of the weights, given by the following optimization problem:

$$\mathbf{w}^* = \text{argmin} \sum_{i=1}^{N} ||\mathbf{w}^T \mathbf{x}_i - y_i||_1 + \lambda ||\mathbf{w}||_2^2$$

What properties would you expect this regression algorithm to have? Explain why?

Q13 ANSWER:

<ANSWER HERE>

- Since the L1-norm is used on the error (residual), the learned function should be robust to outlier values of y_i. This is because there is no preference to reduce the larger errors, in contrast to L2 norm error.
- Since the L2-norm is used on the weights, weights will be shrunk towards zero, or large-weights are discouraged. Thus the learned function should not overfit as long as the regularization parameter is large enough.

--- END ---