ECE523: Engineering Applications of Machine Learning and Data Analytics

I acknowledge that this exam is solely my effort. I have done this work by myself. I have not consulted with others about this exam in any way. I have not received outside aid (outside of my own brain) on this exam. I understand that violation of these rules contradicts the class policy on academic integrity.

Name: ______________________________

Signature: ______________________________

Date: ______________________________

Instructions: There are five problems. You have 50 minutes to complete the exam. You may use handwritten notes on both sides of one 8.5” × 11” piece of paper. Use of any other notes, textbooks, or any other form of outside help is strictly forbidden. Partial credit is given for answers that are partially correct. No credit is given for answers that are wrong or illegible. Write neatly.

Problem 1: ____________

Problem 2: ____________

Problem 3: ____________

Problem 4: ____________

Problem 5: ____________

Total: ____________

Problem #1 – Random Questions (10 Points)

Question 1 In k-nearest neighbors (KNN), the classification is achieved by majority vote in the vicinity of a data sample $\mathbf{x}$. Suppose there are two classes, where each class has $n/2$ points overlapped to some extent in a 2-D space (also, n is even to make this problem easier). Describe what happens to the training error (using all available data) when the neighbor size k varies from n to 1.

Question 2 Consider k-fold cross-validation. Let us consider the tradeoffs of larger or smaller k (the number of folds). With a higher number of folds, the estimated error will be, on average higher, lower, about the same, or don't know.

Question 3 Consider the logistic regression classifier with an output of $\sigma(\mathbf{x}) = 1/(1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}))$. We minimize the cross-entropy function plus a term for L_2 regularization on $\mathbf{w}$. Is this still a convex optimization problem? You must prove your result (i.e., saying yes or no will not earn you the points)

Problem #2 – Short Answer (10 Points)

Question You trained a binary classifier model which gives very high accuracy on the training data, but much lower accuracy on validation data. What is happening?

Question Suppose you are dealing with 4 class classification problem and you want to train a SVM model on the data for that you are using One-vs-all method. How many times we need to train our SVM model in such case?

Question What is the naïve Bayes assumption that is sometimes used in classification? Be as formal and specific as possible.

Question You have just been given a ten datasets from your company and they want you to compare your new classification method to what they currently have. Describe a using techniques that we have discussed in class to make this comparison.

Problem #3 – A Gambler's Ruin (10 Points)

[True/False] (1 point): The support vectors in the context of an SVM are the $\mathbf{x}_i \in \mathcal{D}_{\text{train}}$ (i.e., data set) that correspond to $\alpha_i = 0$.

$$\max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\mathsf{T}} \mathbf{x}_j \right\}$$

$$\text{s.t.} \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \text{ and } 0 \leq \alpha_i \leq C$$

[True/False] (1 point): Increasing the term C in a support vector machine will decrease the number of support vectors.

[True/False] (1 point): Parzen windows are one way to estimate the density, $p_n(x)$, but they cannot be used in classification.

[True/False] (1 point): Regularization is one way to prevent overfitting and the reason it is so effective is because the regularization term is data-independent. Therefore, the optimization process will "find" the best way to be resilient against overfitting.

[True/False] (1 point): The signed-rank Wilcoxon test is the best way to show that multiple classifiers are performing better than each other.

[True/False] (1 point): The Friedman test is the best way to show that multiple classifiers are performing better than each other.

[True/False] (1 point): Parzen windows estimate the density of a dataset by growing a volume V until there are k samples that are enclosed on the region $\mathcal{R}$ with volume V.

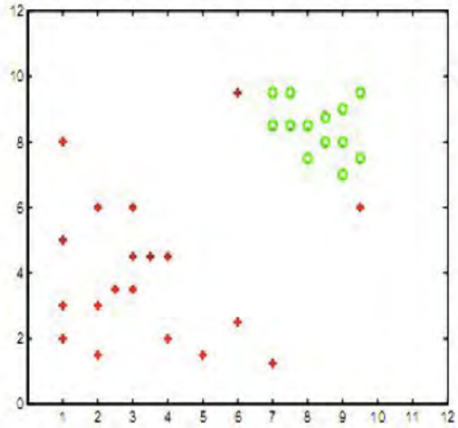
[True/False] (1 point): The testing error of a 1-NN classifier is always zero.

[Accept/Reject] (1 point): "My algorithm is better than yours. Look at the test error rates! (Footnote: reported results for best value of λ, chosen with 10-fold cross validation.)"

[Accept/Reject] (1 point): We did not standardize the features of our data and we did 10-fold cross validation with a k-NN classifier.

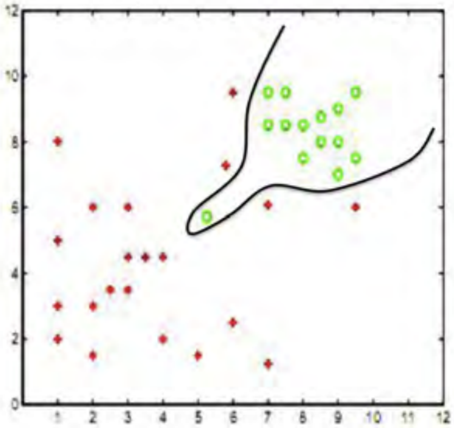
Problem #4 – Support Vector Machines (10 Points)

The original SVM proposed was a linear classifier. In order to make SVM non-linear we map the training data on to a higher dimensional feature space and then use a linear classifier in the that space. This mapping can be done with the help of kernel functions. For this question assume that we are training an SVM with a quadratic kernel - i.e. our kernel function is a polynomial kernel of degree 2. This means the resulting decision boundary in the original feature space may be parabolic in nature. The dataset on which we are training is given below



The slack penalty C will determine the location of the separating parabola. Please answer the following questions qualitatively.

- Where would the decision boundary be for very large values of C? (Remember that we are using a quadratic kernel). Justify your answer in one sentence and then draw the decision boundary in the figure below.
- Where would the decision boundary be for C nearly equal to 0? Justify your answer in one sentence and then draw the decision boundary in the figure below.
- Now suppose we add three more data points as shown in figure below. Now the data are not quadratically separable, therefore we decide to use a degree-5 kernel and find the following decision boundary. Most probably, our SVM suffers from a phenomenon which will cause wrong classification of new data points. Name that phenomenon, and in one sentence, explain what it is.



Problem #4 – Support Vector Machines (10 Points)

Problem #5 – Support Vector Machines for Detection (10 Points)

Let say we want to perform novelty detection with a support vector machine. The concept of performing this task with a hyperplane classifier is possible by reformulating the problem. We have a dataset $\{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\} \subset \mathcal{X}$ and these data samples are considered unlabelled (i.e., there is no y_i label). Also note that $\mathcal{X}$ is a subspace over $\mathbb{R}^d$ where $\mathbf{x}_i$ is d-dimensional. Our goal is to develop a support vector machine to say is a new sample, $\mathbf{x}$, appears to be from the same distribution that generated $\mathcal{X}$ or if it is different $\mathcal{X}'$. The SVM formulation to determine if a sample is an outlier is

$$\arg \min_{\mathbf{w}\in\mathbb{R}^d,\, \zeta\mathbb{R}^d_+,\, \rho\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{1}{\nu n}\sum_{i=1}^n \zeta_i - \rho$$

$$\text{s.t. } \mathbf{w}^\mathsf{T}\mathbf{x}_i \geq \rho - \zeta_i, \quad \zeta_i \geq 0 \quad \forall i \in \{1, \dots, i, \dots, n\}$$

where ζ_i are slack variables, $\nu \in (0, 1)$ is a free parameter and the prediction for outliers is given by

$$f(\mathbf{x}) = \text{sign}\left(\mathbf{w}^\mathsf{T}\mathbf{x} - \rho\right)$$

which will equal 1 if $\mathbf{x}$ is in $\mathcal{X}$ and -1 if it is not. Clearly, the Lag the Langrangian needs to be formed then we can use the Lagrangian to find the detectors parameters. The Lagrangian is given by

$$L(\mathbf{w}, \zeta, \rho, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{1}{\nu n}\sum_{i=1}^n \zeta_i - \rho - \sum_{i=1}^n \alpha_i\left(\mathbf{w}^\mathsf{T}\mathbf{x}_i - \rho + \zeta_i\right) - \sum_{i=1}^n \beta_i\zeta_i$$

where $\alpha_i, \beta_i \geq 0$ are Lagrange multipliers. Using this information

- What is $\mathbf{w}$? How is it different / similar to the support vector machine we discussed in class?
- Are there constraints in α_i? If so what are they?

Problem #5 – Support Vector Machines for Detection (10 Points)