CSE 575 Statistical Machine Learning
Midterm Exam #2
October 25, 2016

1. Personal info:
    - Name:
    - ASU ID#:
2. There should be 10 numbered pages in this exam (including this cover sheet).
3. THIS IS CLOSED BOOK EXAM!
4. If you need more room to work out your answer to a question, use the back of the page and clearly mark on the front of the page if we are to look at what's on the back.
5. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult ones before you have answered the easier ones.
6. You have 75 minutes.
7. Good luck!

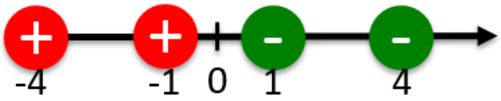
Question	Topic	Max. score	Score
1	SVM and LOOCV	40	
2	K-fold Cross-Validation	20	
3	Loss Function	20	
4	Boosting	20	

1 [40 points] Support Vector Machines and Leave-One-Out-Cross-Validation

Given the following data set in 1-d space, which consists of 2 positive data points at the following coordinates $\{-1, -4\}$ and 2 negative data points at the following coordinates $\{4, 1\}$. Suppose we use a linear SVM (i.e., no kernel), with some large C value, where C is the regularization parameter to control the mis-classification error on the training data set.

1. [4 points] Draw the data set in 1-d space.

   Solution.

2. [5 points] Draw the decision boundary of linear SVM trained on this data set.

   Solution. At the origin.

3. [5 points] In your linear SVM, how many support vectors are there? What are they?

   Solution. 2. -1 and 1.

4. [5 points] What is the leave-one-out-cross-validation (LOOCV) error in your linear SVM? Justify your answer.

   Solution. 0.5

5. [8 points] Now, given another new data set in 1-d space, which consists of 3 positive data points at the following coordinates $\{-0.8, -1, -4\}$ and 2 negative data points at the following coordinates $\{4, 1\}$. Suppose we use a linear SVM (i.e., no kernel), with some large C value. How many support vectors are there in your linear SVM (4 points)? What is the leave-one-out-cross-validation (LOOCV) error in your linear SVM (4 points) ? Justify your answer.

   Solution. 2 support vectors. LOOCV: 0.2.

6. [8 points] Draw the data set from the last question (question #5). In this figure, circle the points such that after removing that point (example) from the training set and retraining SVM, we would get a different decision boundary than training on the full sample.

   Solution. 2 points. -0.8 and 1.

7. [8 points] Draw the data set from question #5. Suppose that instead of SVM, we use regularized logistic regression to learn the classifier. That is,

$$(w, b) = \arg \min_{w \in \mathbb{R}^2, b \in \mathbb{R}} \frac{\|w\|^2}{2} - \sum_i \mathbb{1}[y_i = 0] \ln \frac{1}{1 + e^{(w \cdot x_i + b)}} + \mathbb{1}[y_i = 1] \ln \frac{e^{(w \cdot x_i + b)}}{1 + e^{(w \cdot x_i + b)}}.$$

In this figure, circle the points such that after removing that point (example) from the training set and running regularized logistic regression, we would get a different decision boundary than training with regularized logistic regression on the full sample.

Solution. All the points.

2 [20 points] K-fold Cross-Validation

1 [5 points] Given a dataset with $10,000$ data points, we perform a 5-fold cross-validation, how many data points do we set aside for evaluation (i.e., to calculate the test error) at each iteration?

Solution. $10,000/5 = 2,000$

2 [5 points] Given a dataset with $2,000$ data points, we perform a 100-fold cross-validation, how many data points do we set aside for evaluation (i.e., to calculate the test error) at each iteration?

Solution. $2,000/100 = 20$

3 [5 points] Given a dataset with $1,000$ data points, we perform a k-fold cross-validation, and at each iteration, we set aside 1 single data point for evaluation (i.e., to calculate the test error). What is the k value? Justify your answer.

Solution. $k = 1,000$

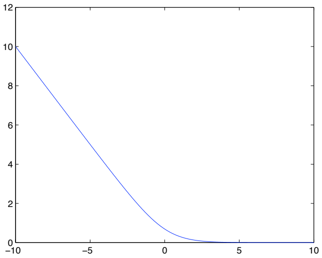
4 [5 points] **True or False**. Given a dataset with N data points, where N is an even number. If we perform a k-fold cross-validation, then at each iteration, we will set aside **at most** $N/2$ data points for evaluation (i.e., to calculate the test error). Justify your answer.

Solution. True.

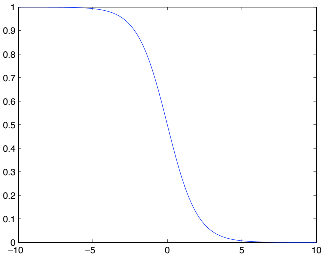
3 [20 points] Loss Function

Generally speaking, a classifier can be written as $H(x) = \text{sgn}(F(x))$, where $x \in \mathbb{R}^d$, i.e., x is in the d-dimensional real space, $\text{sgn}(\cdot)$ is the sign function defined as follows: $\text{sgn}(\cdot)$ is 1 if the input is bigger than 0, -1 if the input is less than or equal to 0. Therefore, $F(x) : \mathbb{R}^d \to \mathbb{R}$ and $H(x) : \mathbb{R}^d \to \{-1, 1\}$. To obtain the parameters in $F(x)$, we need to minimize the loss function averaged over the training set: $\frac{1}{N}\sum_i^N L(y_i F(x_i))$, where $y_i \in \{-1, 1\}$, and $L(\cdot)$ is a function of $y_i F(x_i)$. For example, for linear classifiers, $F(x_i) = w_0 + \sum_{j=1}^d w^j x_i^j$, and $y_i F(x_i) = y_i(w_0 + \sum_{j=1}^d w^j x_i^j)$, where x_i^j is the j^{th} feature in the i^{th} sample.

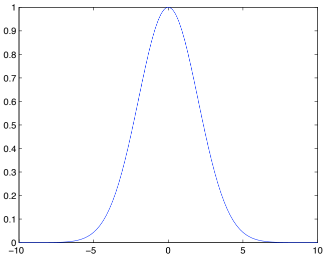
1. [10 points] Which loss functions below are appropriate to use in classification? In general, what conditions does L have to satisfy in order to be an appropriate loss function? The horizontal axis is $yF(x)$, and the vertical axis is $L(yF(x))$.
Hint: The answer includes multiple loss functions.



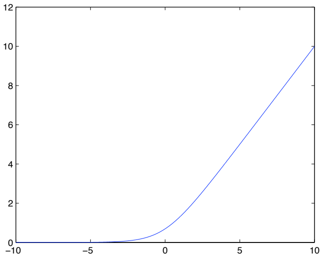
(a)



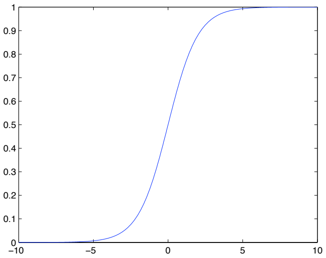
(b)



(c)



(d)



(e)

Solution. (a) and (b) are appropriate to use in classification. In (c), there is very little penalty for extremely misclassified samples, which correspond to very negative $yF(x)$. In (d) and (e), correctly classified samples are penalized, whereas misclassified samples are not. In general, L should approximate the 0-1 loss, and it should be a non-increasing function of $yF(x)$.

2. [10 points] Of the above loss functions appropriate to use in classification, which one is the most robust to outliers? Justify your answer.

   Hint: You can think of an outlier as an example/label pair (x, y) in the training set such that according to the underlying distribution, $P(y|x) \ll 1 - P(y|x)$, i.e., the probability of observing the opposite label is much higher. Outliers are some times due to mistakes. For example, when an example $(x, 1)$ is being added to the training set, someone mistakenly adds $(x, -1)$ instead.

   Solution. (b) is more robust to outliers. For outliers, $yF(x)$ is often very negative. In (a), outliers are heavily penalized. So the resulting classifier is largely affected by the outliers. On the other hand, in (b), the loss of outliers is bounded. So the resulting classifier is less affected by the outliers, and thus more robust.

4 [20 points] Boosting

Consider the AdaBoost algorithm you saw in class. In this question we will try to analyze its training error.

1. [10 points] Given a set of m examples, (x_i, y_i) ($y_i \in \{-1, 1\}$ is the class label of x_i), $i = 1, \ldots, m$, let $h_t(x)$ be the weak classifier obtained at step t, and let α_t be its weight. Recall that the final classifier is

$$H(x) = \text{sign}(f(x)), \text{ where } f(x) = \sum_{t=1}^{T} \alpha_t h_t(x).$$

Show that the training error of the final classifier can be bounded from above by an exponential loss function:

$$\frac{1}{m} \sum_{i=1}^{m} I(H(x_i) \neq y_i) \leq \frac{1}{m} \sum_{i=1}^{m} \exp(-f(x_i) y_i),$$

where $I(a = b)$ is the indicator function. It is equal to 1 if $a = b$, and 0 otherwise.
Hint: $e^{-x} \geq 1 \Leftrightarrow x \leq 0$.

Solution. Consider two outcomes of $I(H(x_i) \neq y_i)$. If $H(x_i) \neq y_i$, then $I(H(x_i) \neq y_i) = 1 < \exp(-f(x_i)y_i)$; otherwise, $I(H(x_i) \neq y_i) = 0 < \exp(-f(x_i)y_i)$.

2. [10 points] In boosting, would you stop the iteration if the following happens? Justify your answer with at most two sentences each question.

   - [5 points] The error rate of the combined classifier on the original training data is 0.

     Solution. No. The test error might continue to decrease even after the error rate of the combined classifier on the original training data reaches 0.

   - [5 points] The error rate of the current weak classifier on the weighted training data is 0.

     Solution. Yes. If the error rate of the current weak classifier on the weighted training data is 0, in the next iteration, all the training data will have weight 0.