CSE 250B: Machine learning **Winter 2019**

Homework 4 — Multiclass, kernels, and theory of generalization

This homework is not meant to be turned in. Try it on your own, and compare your answers to the solution set that will be released on Tuesday March 5.

1. A linear predictor is used to solve a classification problem with three classes. The data is two-dimensional and the linear functions for each class are:

   - Class 1: $w_1 = (1, 1)$, $b_1 = 0$
   - Class 2: $w_2 = (1, 0)$, $b_2 = 1$
   - Class 3: $w_3 = (0, 1)$, $b_3 = -1$

   Draw the resulting decision boundary and clearly mark the region corresponding to each class.

2. *Multiclass Perceptron.* Implement the multiclass Perceptron algorithm from class.

   (a) Load in the data set `data0.txt`. This has 2-d data in four classes (coded as 0,1,2,3). Each row consists of three numbers: the two coordinates of the data points and the label.

   (b) Run the multiclass Perceptron algorithm to learn a classifier. Create a plot that shows all the data points (with different colors and shapes for different labels) as well as the decision regions.

3. *Kernel Perceptron.* Implement the kernel Perceptron algorithm, with the quadratic and RBF kernels.

   (a) The data sets `data1.txt` and `data2.txt` contain 2-d data with two classes (coded as -1 and 1). Each row has three numbers: the two coordinates of the data points and the label.

   (b) Run the kernel Perceptron with quadratic kernel on these two data sets. In each case, show a plot that contains all the data points (with different colors and shapes for different labels) as well as the decision region.

   (c) Repeat for the RBF kernel. Show the results for two different settings of the scale parameter σ.

4. *Multiclass kernel SVM.* In this problem, we'll use support vector machines to classify the MNIST data set of handwritten digits.

   (a) Load in the MNIST data: a training set of 60,000 points and a separate test set of 10,000 points.

   (b) Learn a linear SVM classifier using `sklearn.svm.LinearSVC`. You will need to see `loss='hinge'`. Try different values of the tradeoff parameter: $C = 0.01, 0.1, 1.0, 10.0, 100.0$. In each case, report the training error and test error. Is this data linearly separable?

   (c) Now try kernel SVM with a quadratic kernel. You can do this with `sklearn.svm.SVC`, setting `kernel='poly'` and `degree=2`. Just try the single setting $C = 1.0$. Report the training error, the test error, and the number of support vectors.

5. *Pointwise product of positive semidefinite matrices.* In this problem, we'll see that if M, N are $d \times d$ symmetric positive semidefinite matrices, then so is their *pointwise* product, Q:

$$Q(i, j) = M(i, j)N(i, j).$$

We will make use of the following facts: any covariance matrix is positive semidefinite; and any positive semidefinite matrix is the covariance matrix of some vector-valued random variable.

(a) Suppose random variable $X \in \mathbb{R}^d$ has a Gaussian distribution with mean zero and covariance M. Likewise, suppose $Y \in \mathbb{R}^d$ is Gaussian with mean zero and covariance N, and is independent of X. Define the pointwise product of X and Y to be $Z \in \mathbb{R}^d$:

$$Z_i = X_i Y_i.$$

What are the mean and covariance of Z?

(b) Conclude that Q is positive semidefinite.

6. *Closure properties of kernels.* Suppose k_1 and k_2 are two kernel functions on a space $\mathcal{X}$.

(a) Show that $k(x, x') = \alpha_1 k_1(x, x') + \alpha_2 k_2(x, x')$ is also a kernel function, for any $\alpha_1, \alpha_2 \geq 0$.

(b) Show that $k(x, x') = k_1(x, x')k_2(x, x')$ is also a kernel function. *Hint:* Remember our earlier result about the pointwise product of PSD matrices.

7. Show that $k(x, z) = (x \cdot z)^2$ is a kernel function (on $\mathbb{R}^d$) by exhibiting an explicit embedding $\Phi(\cdot)$ such that $k(x, z) = \Phi(x) \cdot \Phi(z)$. *Hint:* It might help to start by looking at the special cases $d = 1, 2, 3$.

8. Let $\mathcal{X} = \{0, 1\}^d$. The class $\mathcal{H}$ of *monotone disjunctions* consists of classifiers h that are given by a disjunction (logical OR) of some subset of the d features. For instance, the classifier

$$h(x) = x_1 \vee x_3 \vee x_8$$

assigns label 1 to points $x \in \mathcal{X}$ for which any of the features x_1, x_3, x_8 are set; and assigns label 0 otherwise. Suppose we obtain a training set of n points, drawn i.i.d. from an unknown underlying distribution, and we find a monotone disjunction $h \in \mathcal{H}$ that is correct on all n points. We would like to give a bound on the true error of h.

(a) What is $|\mathcal{H}|$? Your answer should be a function of d.

(b) Give a bound on the true error of h that holds with probability at least $1 - \delta$ over the choice of training data.

(c) What bound could you give if instead we looked at the smaller class $\mathcal{H}_k \subset \mathcal{H}$ of *k-sparse monotone disjunctions*: that is, monotone disjunctions consisting of at least 1 and at most k variables?

9. *Estimating the bias of a coin.* A coin of bias 3/4 is tossed 300 times and an empirical estimate $\widehat{p}$ of the bias is obtained. Use the central limit theorem to come up with an interval in which $\widehat{p}$ will lie, with 95% probability.

10. Determine the VC dimension of the following concept classes. Justify your answers.

(a) *Intervals on the line.* $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ where $h_{a,b}(x) = 1(a \leq x \leq b)$.

(b) *Axis-aligned rectangles in the plane.* Each $h \in \mathcal{H}$ is given by an axis-aligned rectangle in $\mathbb{R}^2$, where points inside the rectangle are labeled 1, and points outside are labeled 0.