EE-UY 4423: Introduction to Machine Learning
Midterm 2, Fall 2016

Answer all THREE questions. Exam is closed book. No electronic aids. But, you are permitted a limited number of cheat sheets. Part marks are given. If you do not remember a particular python command or its syntax, use pseudo-code and state what syntax you are assuming.

Best of luck!

1. A data scientist is hired by a conservative political candidate to predict who will donate money. The data scientist decides to use two predictors for each possible donor:

   - x_1 = their income (in thousands of dollars), and
   - x_2 = the number of conservative websites they follow on Facebook.

   To train the model, the scientist tries to solicit donations from a randomly selected subset of people and records who donates or not. She obtains the following data:

Income (thousands \$), x_{i1}	30	50	70	80	100
Num websites, x_{i2}	0	1	1	2	1
Donate (1=yes or 0=no), y_i	0	1	0	1	1

   (a) Draw a scatter plot of the data labeling the two classes with different markers.

   (b) Find a linear classifier that makes at most one error on the training data. The classifier should be of the form,

$$\hat{y}_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0, \end{cases} \quad z_i = \mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b.$$

   What is the weight vector $\mathbf{w}$ and bias b in your classifier?

   (c) Now consider a logistic model of the form,

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1+e^{z_i}}, \quad z_i = \mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b.$$

   Using $\mathbf{w}$ and b from the previous part, which sample i is the *least* likely (i.e. $P(y_i|\mathbf{x}_i)$ is the smallest). If you do the calculations correctly, you should not need a calculator.

   (d) Now consider a new set of parameters

$$\mathbf{w}' = \alpha\mathbf{w}, \quad b' = \alpha b,$$

   where $\alpha > 0$ is a positive scalar. Would using the new parameters change the values $\hat{y}$ in part (b)? Would they change the likelihoods $P(y_i|\mathbf{x}_i)$ in part (c)? If they do not change, state why. If they do change, qualitatively describe the change as a function of α.

   (e) Complete the following python function to generate a vector of random labels `y`, where `y[i]` uses the data record in `X[i,:]`, weight vector `w` and bias `b`.

```
import numpy as np
def gen_rand(X,w,b):
    ...
    return y
```

2. Consider the data set with scalar features x_i and binary class labels $y_i = \pm 1$.

x_i	0	1	3	4	6
y_i	1	-1	1	1	1

A support vector classifier is of the form

$$\hat{y} = \begin{cases} 1 & z > 0 \\ -1 & z < 0, \end{cases} \quad z = \sum_i \alpha_i y_i K(x_i, x),$$

where $K(x, x')$ is the radial basis function, $K(x, x') = e^{-\gamma(x-x')^2}$, and $\gamma > 0$ is a parameters and the dual coefficients are $\boldsymbol{\alpha} = [1, 1, 1, 0, 0]$

(a) Describe a linear classifier for this data. Is the data above linearly separable?

(b) Given the values of $\boldsymbol{\alpha}$ above, which are the support vectors?

(c) Suppose that $\gamma \gg 1$. Approximately draw z vs. x.
Hint: If γ is large, then

$$e^{-\gamma(x-x_i)^2} \gg e^{-\gamma(x-x_j)^2} \text{ if } |x - x_i| < |x - x_j|.$$

(d) For what values of x is $\hat{y} = 1$?

(e) Complete the following python code to output a vector `yhat` of the predicted values of an SVM classifier at a vector of test values `x`. The SVM uses training data `xtr,ytr`, a radial basis function scale factor `gam` and coefficient `alpha`. Assume that the data is scalar so that `x` and `xtr` are a `numpy` 1D arrays. For full credit, vectorize all operations and write the code without any for loops.

```
import numpy as np
def rbf_predict(xtr,ytr,x,gam,alpha):
   ...
   return yhat
```

Hint: For the most efficient implementation, use python broadcasting to first create a matrix `D` with components `D[i,j]=x[i]-xtr[j]`.

3. An audio engineer wants to design a speech classification system using a neural network. A speaker says one of $K = 10$ words and the system is suppose to determine which word was spoken. For training data, the engineer collects $N = 50000$ samples $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$ where each $\mathbf{x}_i$ is a vector of 120 features of the sound and $y_i = 1, \ldots, K$ is the word label. She then tries to fit a neural network of the form,

$$\mathbf{z}_i^{\rm H} = \mathbf{W}^{\rm H}\mathbf{x}_i + \mathbf{b}^{\rm H}, \quad \mathbf{u}_i^{\rm H} = g_{\rm act}(\mathbf{z}_i^{\rm H}),$$
$$\mathbf{z}_i^{\rm O} = \mathbf{W}^{\rm O}\mathbf{u}_i^{\rm H} + \mathbf{b}^{\rm O}, \quad \hat{y}_i = g_{\rm out}(\mathbf{z}_i^{\rm O}),$$

where the activation function is a sigmoid,

$$u_{ij}^{\rm H} = g_{\rm act}(z_{ij}^{\rm H}) = \frac{1}{1 + z_{ij}^{\rm H}},$$

and the output is the argmax,

$$\hat{y}_i = g_{\rm out}(\mathbf{z}_i^{\rm O}) = \arg\max_{k=1,\ldots,K} z_{ik}^{\rm O}.$$

(a) Suppose the network uses $N_h = 50$ hidden units. What are the dimensions of the parameters $\mathbf{W}^{\rm H}$, $\mathbf{b}^{\rm H}$, $\mathbf{W}^{\rm O}$, $\mathbf{b}^{\rm O}$.

(b) Given a set of parameters θ, what is a possible loss function $L(\theta)$ that can be used for training the data?

(c) Draw the computation graph showing the mapping from the inputs $\mathbf{x}_i$ and the parameters θ to the loss function, L. Indicate which terms are data and parameters.

(d) Suppose that in backpropagation, you have already computed the derivative, $\partial L/\partial \mathbf{u}_i^{\rm H}$. Write an expression to compute $\partial L/\partial \mathbf{z}_i^{\rm H}$. What are the dimensions of $\partial L/\partial \mathbf{u}_i^{\rm H}$ and $\partial L/\partial \mathbf{z}_i^{\rm H}$?

(e) Suppose that you have computed $\partial L/\partial \mathbf{z}_i^{\rm H}$ for all i. Write an expression for the gradients $\nabla_{\mathbf{W}^{\rm H}} L$ and $\nabla_{\mathbf{b}^{\rm H}} L$. What are the dimensions of the gradients?