Exercise 3

Prof. Kontorovich and Dr. Sabato

Submission guidelines, **please read and follow carefully**:

- You may submit the exercise in pairs.
- Submit using the submission system.
- The submission should be a zip file named "ex3.zip".
- The zip file should include **exactly two files in the root - no subdirectories please**.
- The files in the zip file should be:
  1. A file called "answers.pdf" - The answers to the questions, including the graphs.
  2. A file called "softsvmpoly.m" - The Matlab/Octave code for the requested function. Note that you can put several auxiliary functions in this file after the definition of the main function. **Make sure that the single file works in Matlab/Octave before you submit it.**
- **Anywhere in the exercise where Matlab is mentioned, you can use Octave instead**.
- Grading: 1(a): 25 points. 1(b): 5 points. 1(c): 10 points. Q. 2-4: 20 points each.
- For questions use the course Forum, or email `inabd17@gmail.com`.

Question 1. For this question, use the data file `EX3q1_data.mat`, which contains data points $x_i \in \mathbb{R}^2$ and labels $y_i \in \{-1, 1\}$. There are 1000 training examples and 100 test points.

(a) Implement the soft-margin kernel SVM routine described in class, using MATLAB's `quadprog` command. Use the Polynomial kernel. The function should be implemented in the submitted file called "softsvmpoly.m". The first line in the file (the signature of the function) should be:

```
function alpha = softsvmpoly(lambda, k, m, d, Xtrain, Ytrain)
```

The input parameters are:

- `lambda` - the parameter λ of the soft SVM algorithm.
- `k` - the degree of the polynomial kernel.
- m - the size of the training sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$, an integer $m \geq 1$.
- `d` - the number of features in each example in the training sample, and integer $d \geq 1$. So $\mathcal{X} = \mathbb{R}^d$.

- `Xtrain` - a 2-D matrix of size $m \times d$ (note: first number is the number of rows, second is the number of columns). Row i in this matrix is a vector with d coordinates that describes example x_i from the training sample.
- `Ytrain` - a column vector of length m (that is, a matrix of size $m \times 1$). The i's number in this vector is the label y_i from the training sample. You can assume that each label is either -1 or 1. **Important: The labels in the input to soft SVM should be -1 or 1, and not 0 or 1.**

The function returns the variable `alpha`. This is the vector of coefficients found by the algorithm, $\alpha \in \mathbb{R}^m$, a column vector of length `d`.

(b) Perform 10-fold cross-validation to tune λ and k. Try 3 different values of λ, and 3 different values of k, a total of 9 parameter pairs that need to be tried. The values for λ should be $0.01, 0.1, 1$ and the values for k should be $3, 10, 50$. Report the 9 validation error values for each of the pairs (λ, k) as well as the optimal pair (i.e., the one that achieves the lowest validation error) and its performance on the test set.

(c) Choose the best value of λ for $k = 3$ that you found in the previous question. Take the output α that you got from this pair, and use it to calculate the predictor w, which is a vector in the new feature space. Answer the following questions:

- What formula did you use to convert α to w?
- List the coordinates of the vector w.
- Write down the multivariate polynomial in x that is generated by the inner product $\langle w, \psi(x) \rangle$.

Question 2. A researcher wants to know what percent of the time commercials were broadcast on channel Z during 2015. Denote this unknown percent α. The researcher draws 100 points in time in 2015 uniformly at random, and checks the logs of channel Z for each of those times, to see whether there was a commercial broadcast at that time. It turns out that the proportion of times with a commercial out of the times that were tested is $\hat{p}$ (a fraction in $[0, 1]$).

(a) The researcher wants to use Hoeffding's bound to draw from $\hat{p}$ conclusions on the value of α. Explain why it is OK to use Hoeffding's bound here:

i. What are the random variables to use in the bound?
ii. Are they independent? why?
iii. What is the probability of each of these random variables to be 1?

(b) Suppose that $\hat{p} = 22\%$. Use Hoeffding's bound to give upper and lower bounds on α. The upper and lower bounds should hold with a probability of 99%. Explain all your calculations and your use of the bound.

Question 3. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Consider a Gradient Descent algorithm that attempts to minimize the following objective:

$$\text{Minimize}_{w \in \mathbb{R}^d} \lambda \|w\| + \sum_{i=1}^{m} (\langle w, x_i \rangle - y_i)^2.$$

(a) Suppose that Gradient Descent is run on S with a step size η. Calculate the the formula for $w^{(t+1)}$ as a function of $w^{(t)}$ and η. Explain the steps of your derivation.

(b) What would the update step for $w^{(t+1)}$ be in Stochastic Gradient Descent for the same objective?

Question 4. Define the **depth** of a tree as the maximal path length from the root to a leaf. Let $\mathcal{X} = \{0,1\}^d$. Let $\mathcal{H}'_n \subseteq \{0,1\}^{\mathcal{X}}$ be the hypothesis class consisting of decision trees with depth at most n and binary attribute tests of the form "$x(i) = 1$?". (Note: this class is different from $\mathcal{H}_n$, which we discussed in class).

(a) Prove that $|\mathcal{H}'_n| \leq (d+2)^{2^n}$.

(b) Suppose an ERM algorithm for $\mathcal{H}'_4$ is performed on a random sample of size m from $\mathcal{D}$. Suppose $d = 3$. Let $\hat{h}$ be the decision tree that is output by this algorithm. Suppose that $\inf_{h \in \mathcal{H}'_4} \mathrm{err}(h, \mathcal{D}) = 0.1$. Use PAC-learning sample complexity upper bounds to calculate the sample size that guarantees that with a probability of at least 95%, $\mathrm{err}(\hat{h}, \mathcal{D}) \leq 0.3$. Explain what formula you used, and justify all the numbers you use in the formula to get the result.