Homework 6

ESE 402/542

Due December 10, 2021 at 11:59pm

Type or scan your answers as a single PDF file and submit on Gradescope.

Problem 1. *Principal Component Analysis.* Consider the following dataset:

x	y
0	1
1	1
2	1
2	3
3	2
3	3
4	5

(a) Standardize the data and derive the two principal components in sorted order. What is the new transformed dataset using the first principal component?

(b) Repeat the previous analysis but this time do not standardize the original data. Is Principal Component Analysis scale invariant?

Problem 2. *Polynomial Regression.* Load the dataset poly_data.csv. The first column is a vector of inputs x and the second column is a vector of responses y. Suppose we believe it was generated by some polynomial of the inputs with Gaussian error, i.e. for some (unknown) p,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. We would like to recover the true coefficients of the underlying process. A polynomial regression can be estimated by including all powers of x as predictors in the model (see recitation 7 for details). However, the problem is that we don't know what the true value of p is. We will use k-fold cross validation to solve this problem.

(a) Pick a set of polynomial models, i.e. all polynomials of degree 1 to degree 40. Compute the k-fold cross validation error (mean squared error) for each of these models. Report the value of k that you use and plot the cross-validation error as a function of polynomial degree. Which polynomial degree fit the data the best?

(b) Choose the best polynomial model obtained from the previous part, and use to it regress the entire dataset. Report the polynomial coefficients and make a scatter plot of the x_i's and y_i's with your fitted polynomial.

Problem 3. *Bayes Optimal vs. Logistic Regression.* Recall that in classification, we assume that each data point (x_i, y_i) is drawn i.i.d. from a joint distribution P, i.e. $P(X = x, Y = y) = P(Y = y)P(X = x|Y = y))$. In this problem, we will examine a particular distribution on which Logistic Regression is optimal. Suppose that this distribution is supported on $x \in \mathbb{R}, y \in \{-1, +1\}$, and given by:

$$
\begin{aligned}
P(Y = +1) &= P(Y = -1) = 1/2 \\
P(X = x|Y = +1) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-5)^2}{2}} \\
P(X = x|Y = -1) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+5)^2}{2}}
\end{aligned}
$$

(a) Show that the joint data distribution is given by

$$P(X = x, Y = y) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-5y)^2}{2}}$$

(b) Plot (either using code or hand-drawn neatly) the conditional distributions $P(X = x|Y = +1)$ and $P(X = x|Y = -1)$ in a single figure. Note: these are just Gaussian PDFs.

(c) Write the Bayes optimal classifier $h^*(x)$ given the above distribution P and simplify. Hint: you should get a classification rule that classifies x based on whether or not it is greater than a threshold.

(d) Compute the classification error rate of the Bayes optimal classifier, i.e.

$$\Pr_{(x,y)\sim P}(h^*(x) \neq y) = \mathop{\mathbb{E}}_{(x,y)\sim P}[\mathbb{1}_{h^*(x)\neq y}]$$

Hint: Your result should be of the form $1 - \Phi(c)$, where $\Phi(\cdot)$ is the Gaussian CDF.

(e) (Extra Credit) Recall that Logistic Regression assumes that the data distribution is of the form

$$P(Y = +1|X = x) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x}}$$

Show that the distribution given above satisfies this assumption. What values of β_0, β_1 does this correspond to?

Problem 4. (Extra Credit) *k-means is suboptimal.* Recall that the k-means algorithm attempts to minimize the following objective:

$$\min_{c_1,\ldots,c_k} \sum_{i=1}^{n} \|x_i - c(x_i)\|_2^2 \tag{1}$$

where $c(x_i)$ is the closest center to x_i. Show that the k-means algorithm algorithm does not always find the optimal solution of the above objective.

Hint 1: Let OPT denote the optimal objective. For every $t > 1$, show there exists an instance of the above optimization problem for which the k-means algorithm *might* find a solution whose objective value at least t·OPT. In other words, find a set of points $x_1, \ldots, x_n$ for which k-means, with some bad initialization of the centers, will output a set of centers that achieves an objective of t·OPT.

Hint 2: Start with an example of 4 points in a 2-D plane, with 2 clusters. You can then generalize this example to arbitrary p dimensions, n data points, and k clusters.

Problem 5. (Extra Credit) Load the Labeled Faces in the Wild dataset from sklearn. You can load this data as follows:

```
from sklearn.datasets import fetch_lfw_people
faces = fetch_lfw_people(min_faces_per_person=60)
```

For this exercise, we will use PCA on image data, in particular pictures of faces, to extract features.

(a) Perform PCA on the dataset to find the first 150 components. Since this is a large dataset, you should use randomized PCA instead, which can also be found on sklearn. Show the eigenfaces associated with the first 1 through 25 principal components.

(b) Using the first 150 components you found, reconstruct a few faces of your choice and compare them with the original input images.