EE104 final exam

- This is a 24-hour take-home exam. Please turn it in on Gradescope. Be aware that you must turn it in within 24 hours of downloading it. After that, Gradescope will not let you turn it in and we cannot accept it.

- The exam should not be discussed at all until 6/8 after everyone has taken the exam.

- If you have a question, please submit a private question on Ed, or email the staff mailing list. We have tried very hard to make the exam unambiguous and clear, so unless there is a mistake on the exam we’re unlikely to say much.

- We expect your solutions to be legible, neat, and clear. Do not hand in your rough notes, and please try to simplify your solutions as much as you can. We will deduct points from solutions that are technically correct, but much more complicated than they need to be.

- Please check your email during the exam, just in case we need to send out a clarification or other announcement.

- Start each solution on a new page.

- When a problem involves some computation (say, using Julia), we do not want just the final answers. We want a clear discussion and justification of exactly what you did as well as the final numerical result.

- Because this is an exam, you must turn in your code. Include the code in your pdf submission. We reserve the right to deduct points for missing code.

- In the portion of your solutions where you explain the mathematical approach, you cannot refer to Julia operators, such as the backslash operator. (You can, of course, refer to inverses of matrices, or any other standard mathematical constructs.)

- Some of the problems require you to download data or other files. These files can be found at the URL [http://ee104.stanford.edu/final22.html](http://ee104.stanford.edu/final22.html)

- Good luck!
1. Medical costs

The file `medicalinsurance.json` contains the simplified medical record data from a set of individuals. You will be predicting the medical costs based on the provided data.

Each record \( u \) consists of the following details in order:

- age
- sex (0 is male, 1 is female)
- BMI: body mass index
- number of children
- smoker (0 is no, 1 is yes)
- region (1: northeast, 2: northwest, 3: southeast, 4: southwest)

The target variable \( v \) contains the charges billed by health insurance. We will embed via \( x = \phi(u) \), and \( y = v \).

(a) Use a one-hot embedding for region, and embed the other fields using the identity map. Standardize the training set (do not standardize the target, \( y \)). Report the means and standard deviations of each feature column before standardization.

(b) For the test set, standardize using the mean and standard deviation from the corresponding feature columns in the training set. Report the means and standard deviations of each feature column in the test set before and after this transformation.

(c) Add a constant feature to the data set. For 50 values of \( \lambda \) uniformly spaced on a log scale between \( 10^{-1} \) and \( 10^5 \), fit a ridge regression model to the data. Plot the training and test loss versus \( \log(\lambda) \).

(d) What value of \( \lambda \) would you choose?

(e) Report the smallest RMSE test error and the corresponding \( \lambda \).

(f) For the optimal predictor, report the three features which have most influence on healthcare costs (apart from the constant feature).

(g) Plot the prediction \( \hat{y} \) against the true target \( y \) for the test set.

(h) Plot the target \( y \) versus age for the training set. What do you notice? Can you suggest how you might build a predictor to handle this problem?

Solution.

(a) The means and standard deviations are

```
training set before standardization:
field: age mean = 38.76, std = 13.98
field: sex mean = 0.50, std = 0.50
field: bmi mean = 30.63, std = 6.13
field: children mean = 1.11, std = 1.21
field: smoker mean = 0.21, std = 0.41
field: NE mean = 0.24, std = 0.43
field: NW mean = 0.25, std = 0.43
field: SE mean = 0.27, std = 0.45
field: SW mean = 0.23, std = 0.42
```

(b) Before standardization:
test set before standardization:

field: age  mean = 40.98,  std = 14.21
field: sex   mean =  0.47,  std =  0.50
field: bmi   mean =  30.79, std =  5.98
field: children mean =  1.04, std =  1.20
field: smoker mean =  0.18,  std =  0.39
field: NE    mean =  0.25,  std =  0.43
field: NW    mean =  0.21,  std =  0.41
field: SE    mean =  0.26,  std =  0.44
field: SW    mean =  0.28,  std =  0.45

test set after standardization:

field: age  mean =  0.16,  std =  1.02
field: sex   mean =  -0.05,  std =  1.00
field: bmi   mean =  0.03,  std =  0.98
field: children mean =  -0.06, std =  1.00
field: smoker mean =  -0.07,  std =  0.95
field: NE    mean =  0.01,  std =  1.01
field: NW    mean =  -0.10,  std =  0.94
field: SE    mean =  -0.03,  std =  0.99
field: SW    mean =  0.12,  std =  1.07

(c) The training rmse (in blue) and test rmse are below. (Points were also awarded for square loss if the values were right.)

(d) The optimal lambda is $\lambda = 0.1$ in the given range, or $\lambda = 0.029$ outside the given range.
(e) The optimal test rmse is 6216 for $\lambda = 0.029$, and 6253 for $\lambda = 0.1$.
(f) The most important features, along with the corresponding $\theta_i$ values are
(g) The performance plot is

![Performance Plot]

(h) The charges versus age plot is below.

![Charges vs Age Plot]

We observe that there are three different categories of charges; low, medium, and high, at all ages. We might handle this by introducing new features, or by predicting three possible charges instead of one.

2. Titanic survivors
In the file `titanic.json`, you will find data about passengers who travelled on the Titanic. Each record \( u \) contains, in order:

- passenger class
- gender (0: male, 1: female)
- age
- number of siblings or spouses aboard
- number of parents or children aboard
- fare
- the port from which the passenger embarked (1: Cherbourg, 2: Queenstown, 3: Southampton)

The target variable \( v \) specifies whether they survived (1 = died, 2 = survived). We have split the data into train and test sets.

(a) First, we consider a simple predictor based purely on the gender of the passengers. Let \( G(u) \) be

\[
G(u) = \begin{cases} 
\text{survived} & \text{if passenger is female} \\
\text{died} & \text{otherwise}
\end{cases}
\]

Find the confusion matrix on the test set for this predictor.

(b) Next we consider a simple predictor based purely on the passenger class. Let \( G(u) \) be

\[
G(u) = \begin{cases} 
\text{survived} & \text{if class < 3} \\
\text{died} & \text{otherwise}
\end{cases}
\]

Find the confusion matrix on the test set for this predictor.

(c) Which of the two simple predictors has the smaller error rate?

(d) Embed the data as follows:

- Embed the embarked feature as one-hot.
- Add an additional boolean feature `alone`. A person travelling alone (zero siblings/spouses and zero parents/children) should have 1 as the value and 0 otherwise.
- Standardize the training data. Standardize the test data using the mean and standard deviation of the training data.
- Add a constant field.

Using logistic loss and \( L_1 \) regularizer, find the predictor that minimizes the empirical risk on the training set. Use regularization parameter \( \lambda = 0.01 \). Report the confusion matrix on the training and test sets for this predictor.

Solution.

(a) the gender-based predictor has confusion matrix

\[
\begin{array}{cc}
83 & 23 \\
15 & 57
\end{array}
\]

(b) the class-based predictor has confusion matrix

\[
\begin{array}{cc}
72 & 26 \\
26 & 54
\end{array}
\]
(c) The gender-based predictor is better. It makes 38 errors, whereas the class-based predictor makes 52 errors.

(d) On the training set, the confusion matrix is

\[
\begin{bmatrix}
392 & 83 \\
59 & 177 \\
\end{bmatrix}
\]

and on the test set it is

\[
\begin{bmatrix}
86 & 22 \\
12 & 58 \\
\end{bmatrix}
\]

3. **Bank note anomalies**

In this problem, you will be clustering bank notes based on different measured parameters. The file `banknotes.json` contains the features for a 200 bank notes, 100 of which are known to be forgeries. The data fields are

- \(u_1\), the length of the bank note
- \(u_2\), the height of the bank note, measured on the left
- \(u_3\), the height of the bank note, measured on the right
- \(u_4\), the distance of inner frame to the lower border
- \(u_5\), the distance of inner frame to the upper border
- \(u_6\), the length of the diagonal

(a) Standardize the data, and let \(x = \phi(u)\) be the resulting embedded data. With \(K = 2\), fit a \(K\)-means model. Report the final embedded cluster centers.

(b) Report the number of data points assigned to each centroid.

(c) Make a plot of \(x_4\) versus \(x_6\) for the data, coloring each point according to which cluster it belongs to. Plot the corresponding (4th and 6th components of) the cluster centers.

**Solution.**

(a) The cluster centers are

\[
c_1 = \begin{bmatrix}
0.143 \\
-0.644 \\
-0.724 \\
-0.797 \\
-0.645 \\
0.940 \\
\end{bmatrix}
\quad c_2 = \begin{bmatrix}
-0.122 \\
0.548 \\
0.616 \\
0.679 \\
0.549 \\
-0.801 \\
\end{bmatrix}
\]

(b) Centroid 1 has 92 points, centroid 2 has 108 points.

(c) The plot is as follows: