CMSC 25025 / STAT 37601

Machine Learning and Large Scale Data Analysis

HW 1

Out: Thursday, March 29, 2018
Due: Thursday, April 5, 2018 (at 1:30 p.m.)

Please hand in this Homework in 5 files:

1. A pdf of the theoretical homework.
2. A pdf of your jupyter notebook for problem 4.
3. The ipynb file for problem 4.
4. A pdf of your jupyter notebook for problem 5.
5. The ipynb file for problem 5.

Read this article and be prepared to discuss in class.
`www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives.`

1. *Random variables* (10 points)

   (a) Let X have a continuous, strictly increasing cumulative distribution function F. Let $Y = F(X)$. Find the density of Y. Now let $U \sim$ Uniform $(0, 1)$ and let $X = F^{-1}(U)$. Show that $X \sim F$.

   (b) Let $X \sim Exp(\lambda), Y \sim Exp(\mu)$ be independent. Find the probability density function for $Z = X - Y$ and $Z = \min\{X, Y\}$.

   (c) Let $X \sim N(0, 1)$ and let $Y = e^X$. Find $\mathbb{E}(Y)$ and Var(Y).

   (d) Show that for any two random variables X, Y Var$(Y) = \mathbb{E}\,\text{Var}(Y \mid X) + \text{Var}\,\mathbb{E}(Y \mid X)$.

2. *Regression* (10 points)

   In linear regression, the fitted values are defined to be $\widehat{\mathbf{y}} = X\widehat{\beta}$ where $\widehat{\mathbf{y}} = H\mathbf{y}$ and

   $$H = X(X^TX)^{-1}X^T.$$

   assuming $n > d$ and X^TX is nonsingular. The matrix H is called the "hat matrix." Define $\mathcal{L}$ to be the set of vectors that can be obtained as linear combinations of the columns of X, which is an $n \times d$ matrix. Show that the hat matrix satisfies the following properties:

   (a) $\widehat{y} = H\mathbf{y} = X\widehat{\beta}$ are the least squares estimates.

   (b) $HX = X$.

   (c) H is symmetric: $H = H^T$.

(d) H is idempotent: $H^2 = H$.

(e) $\widehat{\mathbf{y}} = H\mathbf{y}$ is the projection of $\mathbf{y}$ onto the column space $\mathcal{L}$.

(f) $\mathrm{rank}(X) = \mathrm{tr}(H) = d$.

3. *Singular value decomposition* (10 points)

Let $X \in \mathbb{R}^{m\times n}$ have $\mathrm{rank}(X) = r$ and let $X = U\Sigma V^\top$ be the SVD of A where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m\times n}$ is the diagonal matrix of singular values, with $\Sigma_{ii} = \sigma_i,\ i = 1, ..., r$.

(a) Show that the columns of U are eigenvectors of $XX^\top$ and the columns of V are eigenvectors of $X^\top X$. Determine what are the corresponding eigenvalues in terms of $\sigma_1, ..., \sigma_r$.

(b) If $u_1, ..., u_m$ and $v_1, ..., v_n$ are the columns of U and V, show that $Xv_i = \sigma_i u_i$ and $X^\top u_i = \sigma v_i$.

(c) Express the Frobenius norm $\|X\|_F = \sqrt{\sum_{ij} X_{ij}^2}$ in terms of the singular values σ_i.

(d) Express $|\det(X)|$ in terms of the singular values σ_i (Hint: you will need to use orthogonality to express $|\det(U)|$ and $|\det(V)|$).

(e) Assuming $X^\top X$ is invertible, express the hat matrix $H = X(X^\top X)^{-1}X^\top$ of linear regression in terms of the SVD.

(f) Let $\Sigma^{(k)}, 1 \le k \le r$ be the diagonal matrix $\Sigma_{ii}^{(k)} = \sigma_i$ for $i = 1, ..., k$ (and zeros in the remaining entries). The matrix $U\Sigma^{(k)}V^\top$ is known as the rank-k approximation for X. Express the least square regression estimate obtained using the rank-k approximation instead of X.

4. *Self-fulfilling prophecies* (10 points)

We want to show what kind of issues can arise from clustering when it is applied to real people and affects their choices.

(a) Simulate 1000 points from a bivariate normal distribution $N(\mu, \Sigma)$ with

$$\mu = 0, \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Plot the data. Pretend these two variables represent the features of people an app is using to predict their preference between 3 choices. The features were extracted from data the app developers got from facebook.

(b) Using the python package `sklearn.cluster.KMeans` fit a 3 cluster model to the data. Plot the three cluster centers and color the data points according to the clusters they are assigned. This is the assignment the app has chosen for this population.

(c) Now modify each point in the data to move 1% closer to its assigned cluster center. $x_i = .99 * x_i + .01c_i$, where c_i is the cluster center assigned to the i'th data point. This corresponds to a tiny indirect effect of the choice of cluster on the features of the people in the sample. Now repeat the clustering on the modified data.

(d) Imagine the app repeats the clustering analysis every week based on the modified data. Repeat this process 50 times. Plot the original data cloud, and the final data cloud you obtained side by side. Describe what has happened to your original population of diverse individuals after a year (50 weeks).

5. *Presidential logorrhea* (60 points)

In this problem, you will analyze the lengths of the State of the Union addresses.

(a) The transcripts of all State of the Union addresses are in
`/project/cmsc25025/sou/speeches.pkl`.
You can load them using the pickle package in python as follows:

```
import pickle
f=open('speeches.pkl','r')
speeches=pickle.load(f)
```

Write Python code that parses each SOU address, finding end-of-sentence markers. Don't worry about being too precise about sentence boundaries—as a first approximation, you could find words ending in a period. (But what about "Mr."?)

(b) For each year, compute the number of sentences in the address, and the mean sentence length in words for that year. Plot these data and two linear regressions, one plot for the number of sentences by year, another for the average sentence length by year. Note that the definition of "word" and "sentence" is imprecise. You can experiment with different parsing rules, and see if the results change qualitatively. Describe the trends that you see, and give some explanation for them. You should compute the linear regressions directly—for example, you may use the linear algebra routine `numpy.linalg.solve` but do not use a package that computes the regression.

(c) Now, compute two regressions of the total number of words in a SOU versus year–one for the years 1790 to 1912, another for the years 1913 to the present. What trends do you see? Lookup the history of the State of the Union addresses (for example on Wikipedia) to explain the regressions.

(d) Which President has the longest sentences on average? Which has the shortest sentences? Compute the median, 25% and 75% quantiles across all Presidents. What was the longest and shortest sentence ever spoken (or written) in a SOU?