Assignment 1 V1 - ANLY 601

Due date: January 30th 2019

January 09 2020

Instructions

- Each question is worth a point
- Completing the interview questions are bonus and worth 2 points, each and can be done for extra credit
- If you collaborate with colleagues include their names as *collaborators*
  - You will not be penalized for working together but will be penalized for copying
- If you used references include them as **Bibliography** following the APA style
  - Example: Casella, G., & Berger, R. L. (2002). *Statistical inference (Vol. 2)*. Pacific Grove, CA: Duxbury.
  - You can find a style guide here: http://www.easybib.com/reference/guide/apa/book
- **Assignments need to be typeset in LaTeX and submitted through github as a .pdf**
  - If you need to include graphs you can:
    1. Include them as pictures in the pdf
    2. Use TikZiT or LaTeXDraw
  - Both RStudio and Jupyter notebook have functionality to integrate LaTeX
  - Coding questions should be submitted as `.ipynb` notebooks or runnable`.R`,`.Rmd` files as appropriate

Tips

- **Start working on this assignment early you will not be able to cram it last minute**
- Interview questions are worth more so if you can't do a couple of the required questions try the interview questions (they are bonus but may be asked in a future interview)
- The goal is to show you understand the concepts, this means:
  - For proofs you know to be clear and detailed
  - For mechanical exercises you can skip a couple of steps if the direction is clear
- Try working on problems individually then together as a group
  - The goal is to internalize the concepts so copying doesn't help you

Fundamentals and Review

Exercise 1 (Likelihood Estimation):

1. What is the maximum likelihood estimate for θ when $X_i \sim Geometric(\theta)$?
2. What is the maximum likelihood estimate for a and b when $X_i \sim Unif(a, b)$?

Exercise 2 (Loss Functions):

1. Show that squared error loss (L2 loss) is equivalent to the negative log likelihood of a $Y \sim N(\mu, \sigma^2)$ where σ is known
2. Show that the mean absolute error (L1 loss) is equivalent to the negative log likelihood of a $Y \sim LaPlace(\theta)$

Exercise 3 (Decision Rules):

Suppose that X has mean μ and variance $\sigma^2 < \infty$ show that

1. Show that the mean is optimal decision rule for the mean squared error when the decision rule is unbiased
2. Show the median is the optimal decision rule for the mean absolute error *Hint: do this by minimzing R, see Casella and Berger*

Exercise 4 (Convexity):

Suppose $Y \sim Bernoulli(p)$ where $p = 1/(1 + exp(-\beta x))$ For a fixed x show that:

1. The cross entropy loss $L(y, p) = -(y \log(p) + (1 - y) \log(1 - p))$ is convex with respect to β.
2. The mean squared error loss $L(y, p) = (y - p)^2$ is not convex in β *Hint: use the defintion of convexity*

Exercise 5 (Decision Boundaries):

Suppose $f_\theta(x) = 1/(1 + exp(-\beta x))$ such that $\beta \cdot x = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$

1. If $\theta = 0$ then what is $f_\theta(0) =?$ such that:

$$\begin{cases} \text{Class A} & f_\theta(0) >? \\ \text{Class B} & f_\theta(0) <? \\ \text{Indeterminate} & f_\theta(0) =? \end{cases}$$

   i.e. what is the decision threshold for $f_\theta(0)$ to classify a point is either A or B? What happens if you increase $\beta_0 = 100$? *Hint: plot $f_\theta(x)$ for the simple case $n = 1$*

2. Show that $\theta \cdot \mathbf{x} = \theta_0 + \theta_1 x$ is a linear seperating hyperplane. ($\theta \cdot \mathbf{x} = \theta_0 + \theta_1 x$ is also known as the **linear discriminant**). Show this by taking the logit of $f_\theta(x)$.

Parametric learning

Exercise 6 (Sufficient Statistic)

Suppose $\{X_i\}_{i=1}^{n} \sim N(\mu, \sigma^2); \sigma^2 < \infty$ and σ^2 is known. Show that the sample mean $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ.

Exercise 7 (Ancilliarity)

Choose one:

1. Let $\{X_i\}_{i=1}^n$ be independent and identically distributed observations from a location paramter family with cumulative distribution function $F(x-\theta), -\infty < \theta < \infty$. Show that range of the distribution of $R = \max_i(X_i) - \min_i(X_i)$ does not depend on the parameter θ.) *Hint: Use the facts that* $X_1 = Z_1 + \theta, ..., X_n = Z_n + \theta$ *and* $\min_i(X_i) = \min_i(Z_i + \theta), \max_i(X_i) = \max_i(Z_i + \theta)$, *where* $\{Z_i\}_{i=1}^n$ *are independent and identically distributed observations from* $F(x)$.
2. Show that for $X_{i}{}_{i=1}^n \sim N(0, \sigma^2); \sigma^2 < \infty$ that $\sum_{i=1}^{n-1} \frac{X_i}{X_n} \sim Cauchy(0, n-1)$. *Hint: Show that* $X_i + X_j \sim Cauchy(0, 2\sigma^2)$, *then use ancilliarty (you will need to show this in more a general case).*

Exercise 8 (Completeness)

Show that $N(\mu, \mu^2)$ has a sufficient statistic but is not complete. *Hint: find a linear combination that is not trivially 0 for* $g(T)$

Exercise 9 (Regular exponential family):

Choose one:

1. Show that the *Poisson* distribution is part of the regular exponential family
2. Show that the *GammaDdistribution*is part of the regular exponential family
3. Show that the *multivariate normal* is part of the regular exponential family

Exercise 10 (Regular exponential family)

Show that for the regular exponential family with canonic form that

$$Cov_\eta(T_i(\mathbf{X}), T_j(\mathbf{X})) = \frac{\partial B(\eta)}{\partial \eta_i \partial \eta_j}; i, j \in \{1, 2, ..., n\}$$

Hint: the canonic exponential form is related to the MGF

Exercise 11 (Delta Method)

Suppose we want to estimate the variance of the Bernoulli distribtion $\tau(p) = p(1-p)$ the MLE of this variance is given by $\hat{\tau} = \hat{p}(1-\hat{p})$ where $\hat{p} = \bar{X}$. Using the Delta method find the approximate distribution $\hat{\tau}$.

Information Theory

Exercise 12 (Joint Entropy)

Let $X \in \{0, 1, 2\}$ and $Y \in \{0, 1\}$ be random variables such that their joint distribution is defined as:

$Y \downarrow \| X \rightarrow$	0	1	2
0	1/4	1/12	1/6
1	1/12	1/4	1/6

1. Compute the joint entropy $H(X, Y)$ of X and Y
2. Find the mariginal distribution of X and the conditional entropy $H(Y|X)$
3. Verify the entropy results above by using the chain rules that relates $H(X, Y)$ to $H(X)$ and $H(Y|X)$

Exercise 12 (Differential Entropy)

Find the differential entropy (this is the continuous version of entropy) of a multivariate normal distribution. Use the trick trace$(\mathbf{x}^T\Sigma^{-1}\mathbf{x}) = $ trace$(\Sigma^{-1}\mathbf{x}\mathbf{x}^T)$

Interview questions (Extra Credit)

Relating Ratio of Normals to the Logistic Function - Linear Discrminant Analysis

Suppose that $X \sim MVN(\mu_i, \Sigma) : i = A, B$. Then show that

$$\frac{\Pr(\mathbf{x}|A)}{\Pr(\mathbf{x}|B)} = \frac{1}{1 + exp(-(\boldsymbol{\omega} \cdot \mathbf{x} + \omega_0))}$$

for some $\boldsymbol{\omega}, \omega_0$. Find the values of $\boldsymbol{\omega}, \omega_0$. The resulting function $\phi(\boldsymbol{w}, \mathbf{x}) = (\boldsymbol{\omega} \cdot \mathbf{x} + \omega_0)$ is known as the linear discriminant. What happens when you $\Sigma_A \neq \Sigma_B$

Note that this amounts to looking at the likelihood ratio of two normals and is a form of discrimant analysis. Discriminant analysis is when we have a "discrimnator" that maps from continous space to categorical.

Application of Sufficency

Suppose we were constructing the running average with no buffer in a stream of data.

1. What would be the minimal about of information required to reconstruct the set of data assuming it came from a $Normal(\mu, \sigma^2)$ data set?
2. What is the complete and sufficient statistic for the distribution?

Huffman coding and probability trees

Pinterest is a company that allows you to pin photos that you want to save or share with other users based on their interests. They need their search to be extremely fast. A common approach for search is to build a tree that index searches according to their probability. We're going to explore a simplified version of the approach. Suppose you are given the dictionary made of the following set of words $W = \{cat, dog, shelf, paper, runner, geometric, vase\}$ with the following probabilities

$$\Pr(X = x) = \begin{cases} 1/20 & x = \text{dog} \\ 2/20 & x = \text{cat} \\ 3/20 & x = \text{shelf} \\ 1/20 & x = \text{paper} \\ 6/20 & x = \text{runner} \\ 3/20 & x = \text{geometric} \\ 4/20 & x = \text{vase} \end{cases}$$

and want to encode the various words as a series of bits. For example $dog \to 0, cat \to 1, ...vase \to 0001$. That is you want to generate an encoding for your dictionary. Choose one of the following:

1. If you are familiar with Huffman coding use it to find the optimal tree and encoding. What is your average number of bits needed to encode W
2. If you are unfamiliar with Huffman coding provide a principled way to construct this mapping and estimate the average length of your code.

Research Idea:

Continuing with the example from above. Suppose that instead of W you were given a joint distribution over n set of words i.e. bigrams. How much longer would your average code length be? How would this approach scale as n increases? What about as k increases? Does the underlying distribution of the dictionary matter (look into Zipf's law)? How would you approach scaling the search functionality with other machine learning methods?