DSC 40A - Homework 05

Due: Friday, February 14, 2020

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope on Friday afternoon at 5:00 p.m.

Note: We are collecting mid-quarter feedback! Please consider filling out the following Google Form. Your responses are totally anonymous.

Problem 1.

Let $\vec{y} \in \mathbb{R}^n$, $\mathbb{X}^{n \times (d+1)}$, and $\vec{w} \in \mathbb{R}^{d+1}$.

a) What should be the size of the gradient vector, $\frac{d}{d\vec{w}}[\vec{y}^T X \vec{w}]$? That is, how many entries does it have? You should be able to answer without actually computing the gradient.

b) Compute $\frac{d}{d\vec{w}}[\vec{y}^T X \vec{w}]$ by whatever means you'd like. Simplify your answer as much as possible – it should be some matrix times some vector.

Problem 2.

Beginning with the normal equations, $\vec{w} = (X^\intercal X)^{-1} X^\intercal \vec{y}$, and assuming that $\vec{y}$ is $n \times 1$ and

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

derive the familiar formula

$$w_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Hint: the inverse of a 2×2 matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ is given by $A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$, where $\det(A) = a_{11}a_{22} - a_{12}a_{21}$. This is the only time you'll need to know how to invert a matrix in this class.

Problem 3.

Compute the gradient $\frac{d}{d\vec{x}} \|\vec{x}\|$.

Problem 4.

Suppose you have collected the following data in a survey of data scientists:

Experience	GPA	Salary
5	3.2	85
7	3.7	110
3	3.1	87
9	3.5	105
2	3.2	80

Using least squares, fit a prediction rule of the form $H(\text{experience}, \text{GPA}) = w_0 + w_1 \times \text{experience} + w_2 \times \text{GPA}$. You'll need to solve a system of three equations with three unknowns. You can do this by hand, or you can use `np.linalg.solve`.

Problem 5.

Suppose we collect salaries as well as experience, GPA, and number of internships. To begin, we fit a linear prediction rule $w_0 + w_1 x_1 + w_2 x_2$ to the data using only the experience and GPA as features – we find that our rule has a mean squared error of E_1. Next, we fit a linear prediction rule $w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$ to the same data, but using all three features: experience, GPA, and number of internships. We find that our new line has a mean squared error of E_2. Is it possible for the new mean squared error to be larger than the old? That is, can $E_2 > E_1$? Justify your answer.

Problem 6.

A categorical variable is one which can take only a small number of distinct values. For example, the college that a UCSD student belongs to is a categorical variable: it takes values in the set "Warren", "Muir", "Roosevelt", "Revelle", "Marshall", "Sixth".

We can use categorical variables in regression, but we must first encode the value of a categorical variable as a number. There are two main ways of encoding a categorical variable. The first is to map each possible value to a number; for instance, "Warren" might become 1, "Muir" might become 2, "Roosevelt" becomes 3, and so on. The other approach is called "one-hot encoding". In this situation, we introduce a new feature for every possible value of the categorical variable. To encode someone's college, for instance, we create 6 new features, $x_1, x_2, \ldots, x_6$, one for each college. The feature is one if the student is in that college, and zero otherwise. For instance, supposing that x_1 represents "Warren", x_2 represents "Muir", x_3 represents "Roosevelt", and so on, a student in Muir would be represented as the vector $(0, 1, 0, 0, 0, 0)$.

Suppose we are using least squares regression to find a linear prediction rule for predicting the salary of a UCSD graduate, and we will use their college as a feature. Which method of encoding their college should be used – mapping to a number, or one-hot encoding? Why?