CS 189 Introduction to Machine Learning
Spring 2019 Jonathan Shewchuk

DIS6

1 Linear Regression, Projections and Pseudoinverses

We are given $X \in \mathbb{R}^{n\times d}$ where $n > d$ and $\text{rank}(X) = d$. We are also given a vector $y \in \mathbb{R}^n$. Define the orthogonal projection of y onto range(X) as $P_X(y)$.

(a) Prove that $P_X(y) = \underset{w\in\text{range}(X)}{\arg\min} |y - w|^2$.

Note that in lecture, we learned how to find $\hat{\theta}$ that minimizes the least squares loss $L(\theta) = |y - X\theta|^2$. In other words, we tried to find θ such that $X\theta$ is the vector in the columnspace of X that is closest to our response vector y. Hence, $P_X(y) = X\theta$.

(b) An orthogonal projection is a linear transformation. Hence, we can define $P_X(y) = Py$ for some projection matrix P. Specifically, given $1 \le d \le n$, a matrix $P \in \mathbb{R}^{n\times n}$ is said to be a rank-d orthogonal projection matrix if $\text{rank}(d) = P$, $P = P^\top$ and $P^2 = P$. Prove that P is a rank-d projection matrix if and only if there exists a $U \in \mathbb{R}^{n\times d}$ such that $P = UU^\top$ and $U^\top U = I$

(c) Prove that if P is a rank d projection matrix, then $\text{tr}(P) = d$.

(d) Prove that if $X \in \mathbb{R}^{n\times d}$ and $\text{rank}(X) = d$, then $X(X^\top X)^{-1}X^\top$ is a rank-d orthogonal projection matrix. What is the corresponding matrix U?

For the remainder of the problem set, we no longer assume that X is full rank.

(e) The Singular Value Decomposition theorem states that we can write any matrix X as

$$X = \sum_{i=1}^{\min\{n,d\}} \sigma_i u_i v_i^\top = \sum_{i:\sigma_i>0} \sigma_i u_i v_i^\top$$

where $\sigma_i \ge 0$, and $\{u_i\}$ and $\{v_i\}$ are an orthonormal. Show that

(a) $\{v_i : \sigma_i > 0\}$ are an orthonormal basis for the row space of of X

(b) Similarly, $\{u_i : \sigma_i > 0\}$ are an orthonormal basis for the columnspace of X
Hint: consider $X^\top$.

(f) Define the Moore-Penrose pseudoinverse to be the matrix:

$$X^\dagger = \sum_{i:\sigma_i>0} \sigma_i^{-1} v_i u_i^\top,$$

To what operator does the matrix $X^\dagger X$ correspond? What is $X^\dagger X$ if $\text{rank}(X) = d$? If $\text{rank}(X) = d$ and $n = d$?

2 The Least Norm Solution

Let $X \in \mathbb{R}^{n\times d}$, where $n \geq d$, where $\text{rank}(X)$ is possibly less than d. As in problem 1, we will write the SVD of X as a sum of rank-one terms

$$X = \sum_{i:\sigma_i>0} \sigma_i u_i v_i^\top,$$

In this problem, our goal will to provide an explicit expression for the *least-norm* least-squares estimator, defined as :

$$\widehat{\theta}_{LS,LN} := \arg\min_{\theta}\{|\theta|^2 : \theta \text{ is a minimizer of } |X\theta - y|^2\},$$

where $\theta \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$.

(a) Show that $\widehat{\theta}_{LS,LN}$ is the unique minimizer of $|X\theta - y|^2$ which lies in the rowspace of X. Try not to use the SVD.

(b) Show that $\widehat{\theta}_{LS,LN}$ has the following form:

$$\widehat{\theta}_{LS,LN} = \sum_{i:\sigma_i>0} \frac{1}{\sigma_i} v_i(u_i^\top y), \tag{1}$$

Solve this problem by directly checking that the above expression for $\widehat{\theta}_{LS,LN}$ is in the rowspace of X, and satisfies the necessary optimality condition to be a minimizer of the least-squares objective.

(c) We give another solution to finding a form for $\widehat{\theta}_{LS,LN}$ using the pseudoinverse. Follow these steps:

(1) What is the operator $(X^\top X)^\dagger(X^\top X)$?
Hint: pattern match with the last part of Problem 1, where $X \leftarrow X^\top X$.

(2) Show that $(X^\top X)^\dagger X^\top = X^\dagger$.
*Hint: write everything out as a sum of rank-one terms.*A

(3) Show that any minimizer of the least squares objective satisfies

$$P_X\theta = X^\dagger y,$$

where P_X is the orthogonal projection onto the rowspace of X.

(4) Conclude that

$$\widehat{\theta}_{LS,LN} = X^\dagger y.$$

Verify that this is consistent with your answer to the previous part of the problem.

3 SGD Convergence for Logistic Regression

In this problem, we will prove that gradient descent converges to a unique minimizer of the logistic regression cost function, binary cross-entropy. We will consider the case where we are minimizing this cost function for a single data point. For weights $w \in \mathbb{R}^d$, data $x \in \mathbb{R}^d$, and a label $y \in \{0, 1\}$, the logistic regression cost function is given by

$$J(w) = -y \log s(x \cdot w) - (1 - y) \log(1 - s(x \cdot w))$$

Where $s(\gamma) = 1/(1 + \exp(-\gamma))$ is the logistic function (also called the sigmoid).

(a) To start, write the gradient descent update function $G(w)$, which maps w to the result of a single gradient descent update with learning rate $\epsilon > 0$.

(b) Show that the cost function J has a unique minimizer w^* by proving that J is convex.
Hint: how does this relate to the Hessian, $\nabla_w^2 J$?

(c) Next, show that G is a *contraction*, which means that there is a constant $0 < \rho < 1$ such that, for any $w, w' \in \mathbb{R}^d$, $|G(w) - G(w')| < \rho|w - w'|$. You may assume that the data is bounded, i.e $|x| < c$ for some $c > 0$.
Hint: this is equivalent to showing that the gradient has bounded norm: $\|\nabla_w G(w)\| < \rho$

(d) Finally, calling $w^{(t)}$ the t-th iterate of gradient descent, show that $|w^* - w^{(t)}| < \rho^t|w^* - w^{(0)}|$, so that $\lim_{t\to\infty}|w^* - w^{(t)}| = 0$.