CS 189 Introduction to Machine Learning
Summer 2019 Marc Khoury & Brijen Thananjeyan

DIS8

1 The Ridge Regression Estimator

Recall the ridge estimator for $\lambda > 0$,

$$\widehat{\theta}_\lambda := \arg\min_\theta |X\theta - y|^2 + \lambda|\theta|^2,$$

Let

$$X = U\Sigma V^T = \sum_i \sigma_i u_i v_i^\top$$

be the SVD decomposition of X.

(a) Show that

$$\widehat{\theta}_\lambda = \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \lambda} v_i \langle u_i, y\rangle.$$

(b) Show that

$$|\widehat{\theta}_\lambda|^2 = \sum_{i:\sigma_i>0} \left(\frac{\sigma_i}{\sigma_i^2 + \lambda}\right)^2 \langle u_i, y\rangle^2.$$

(c) Recall the least-norm least squares solution is $\widehat{\theta}_{LN,LS}$ from Problem 2. Show that if $\widehat{\theta}_{LN,LS} = 0$, then $\widehat{\theta}_\lambda = 0$ for all $\lambda > 0$.
Hint: Recall that $\widehat{\theta}_{LN,LS} = \sum_{i:\sigma_i>0} \sigma_i^{-1}\langle u_i, y\rangle v_i$.

(d) Show that if $\widehat{\theta}_{LN,LS} \neq 0$, then the map $\lambda \mapsto |\widehat{\theta}_\lambda|^2$ is strictly decreasing and strictly positive on $(0, \infty)$.

(e) Show that

$$\lim_{\lambda\to 0} \widehat{\theta}_\lambda \to \widehat{\theta}_{LN,LS}.$$

(f) In light of the above, why do you think that people describe the ridge regression as "controlling the complexity" of the solution $\widehat{\theta}_\lambda$?

2 Entropy and Information

In this problem, we try to build intuition as to why entropy of a random variable corresponds to the amount of information that variable transmits. In particular, it determines the number of 0's and 1's needed to "efficiently" encode a random variable.

A coin with bias $b \in (0, 1)$ is flipped until the first head occurs, meaning that each flip gives heads with probability b. Let X denote the number of flips required. Recall that the entropy of a random variable Y is defined as:

$$H(Y) = -\sum_{y} \mathbb{P}(Y = y) \log(\mathbb{P}(Y = y)).$$

(a) Find the entropy $H(X)$. Assuming the logarithm in the definition of entropy has base 2, then the entropy is measured in *bits*.
Hint: The following expressions might be useful:

$$\sum_{n=0}^{\infty} b^n = \frac{1}{1-b}, \quad \sum_{n=1}^{\infty} nb^n = \frac{b}{(1-b)^2}.$$

(b) Let $b = \frac{1}{2}$. Find an "efficient" sequence of yes-no questions of the form, "Is X contained in the set S?", such that X is determined as fast as possible. Compare $H(X)$ to the expected number of asked questions.

3 Decision Trees

Consider constructing a decision tree on data with d features and n training points where each feature is real-valued and each label takes one of m possible values. The splits are two-way, and are chosen to maximize the information gain. We only consider splits that form a linear boundary parallel to one of the axes. We will only consider a standalone decision tree and not a random forest (hence no randomization). Recall the definition of information gain:

$$IG(\textbf{node}) = H(S) - \frac{|S_l|H(S_l) + |S_r|H(S_r)}{|S_l| + |S_r|},$$

where S is set of samples considered at **node**, S_l is the set of samples remaining in the left subtree after **node**, and S_r is the set of samples remaining in the right subtree after **node**.

(a) Prove or give a counter-example: In any path from the root to a leaf, the same feature will never be split on twice. If false, can you modify the conditions of the problem so that this statement is true?

(b) Prove or give a counter-example: The information gain at the root is at least as much as the information gain at any other node.
Hint: Think about the XOR function.

(c) Suppose that a learning algorithm is trying to find a consistent hypothesis when the labels are actually being generated randomly. There are d Boolean features and 1 Boolean label, and examples are drawn uniformly from the set of 2^{d+1} possible examples. Calculate the number of samples required before the probability of finding a contradiction in the data reaches $\frac{1}{2}$.
(A contradiction is reached if two samples with identical features but different labels are drawn.)

(d) Intuitively, how does the bias-variance trade-off relate to the depth of a decision tree?