CS 189 Introduction to Machine Learning
Spring 2021 Jonathan Shewchuk

DIS8

1 The Ridge Regression Estimator

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

$$\widehat{\theta}_\lambda := \arg\min_\theta ||X\theta - y||_2^2 + \lambda||\theta||_2^2,$$

Let

$$X = UDV^T = \sum_i d_i u_i v_i^\top$$

be the SVD decomposition of X. Here U and V are orthogonal matrices where $U^TU = I$ and $V^TV = I$; D is a diagonal matrix.

(a) Show that the optimal weight vector $\widehat{\theta}_\lambda$ can be expressed in the following form:

$$\widehat{\theta}_\lambda = V\Sigma U^\top y$$

where Σ is a diagonal matrix with $\Sigma_{ii} = \frac{d_i}{d_i^2+\lambda}$. Equivalently, we can write $\widehat{\theta}_\lambda$ as

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

Solution: By taking the gradient of the objective, we have that $\widehat{\theta}_\lambda$ has to satisfy

$$X^\top(X\widehat{\theta}_\lambda - y) + \lambda\widehat{\theta}_\lambda = 0,$$

so $\widehat{\theta}_\lambda = (X^\top X + \lambda I)^{-1}X^\top y$. In terms of the SVD of X, this expression is equal to

$$\widehat{\theta}_\lambda = (VDU^\top UDV^\top + \lambda I)^{-1}VDU^\top y = V(D^2 + \lambda I)^{-1}V^\top VDU^T = V\Sigma U^T y,$$

where Σ is a diagonal matrix with $\Sigma_{ii} = \frac{d_i}{d_i^2+\lambda}$. Equivalently, we can write this as

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

(b) Show that

$$\|\widehat{\theta}_\lambda\|_2^2 = \sum_{i:d_i>0} \left(\frac{d_i}{d_i^2+\lambda}\right)^2 (u_i^\top y)^2.$$

Solution: First, we have that

$$\begin{aligned}
\|\widehat{\theta}_\lambda\|_2^2 &= y^\top U\Sigma V^\top V\Sigma U^\top y \\
&= y^\top U\Sigma^2 U^\top y \\
&= \sum_{1\le i\le d} \left(\frac{d_i}{d_i^2+\lambda}\right)^2 \langle u_i, y\rangle^2. \\
&= \sum_{i:d_i>0} \left(\frac{d_i}{d_i^2+\lambda}\right)^2 \langle u_i, y\rangle^2.
\end{aligned}$$

(c) Recall the least-norm least squares solution is $\widehat{\theta}_{LN,LS}$ from DIS6. Show that if $\widehat{\theta}_{LN,LS} = 0$, then $\widehat{\theta}_\lambda = 0$ for all $\lambda > 0$.
Hint: Recall that in DIS6 we showed that $\widehat{\theta}_{LN,LS} = \sum_{i:d_i>0} d_i^{-1}\langle u_i, y\rangle v_i$. This shows that in the case where the least norm least square solution is zero, the ridge regression solution is also zero.

Solution: If the least norm least squares solution is 0, then

$$\sum_{i:d_i>0} d_i^{-1}\langle u_i, y\rangle v_i = 0,$$

which means that $\langle u_i, y\rangle = 0$ for each i where $d_i > 0$, because v_i are linearly independent. Hence, $\widehat{\theta}_\lambda = 0$ by plugging into the formula for $\widehat{\theta}_\lambda = 0$.

(d) Show that if $\widehat{\theta}_{LN,LS} \neq 0$, then the function $f(\lambda) = \|\widehat{\theta}_\lambda\|_2^2$ is strictly decreasing and strictly positive on $(0, +\infty)$.

Solution: If $\widehat{\theta}_\lambda \neq 0$, then at least one of the terms $\langle u_i, y\rangle^2$ is strictly greater than zero. Thus,

$$\|\widehat{\theta}_\lambda\|_2^2 = \sum_{i:d_i>0} \left(\frac{d_i}{d_i^2+\lambda}\right)^2 \langle u_i, y\rangle^2,$$

is a non-trivial nonnegative linear combination of terms $(\frac{d_i}{d_i^2+\lambda})^2$, which are positive and strictly decreasing in λ.

(e) Show that

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

Note that just because the limit of the ridge-regression objective as $\lambda \to 0^+$ is the least squares objective, this does not immediately guarantee that the limit of the ridge solution is the least squares solution.

Solution: Start with the form

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

Since limits commute with sums, we have

$$\lim_{\lambda \to 0^+} \widehat{\theta}_\lambda = \sum_{i=1}^{n} v_i \langle u_i, y \rangle \cdot \left(\lim_{\lambda \to 0^+} \frac{d_i}{d_i^2 + \lambda} \right)$$

Now, we have to consider the cases where $d_i = 0$ and $d_i \geq 0$ (d_i cannot be negative by the properties of SVD).

$$\lim_{\lambda \to 0^+} \frac{d_i}{d_i^2 + \lambda} = \begin{cases} 0 & d_i = 0 \\ d_i^{-1} & d_i > 0 \end{cases}.$$

Thus,

$$\lim_{\lambda \to 0^+} \widehat{\theta}_\lambda = \sum_{i : d_i > 0} d_i^{-1} v_i \langle u_i, y \rangle,$$

which we have shown above is the least norm solution.

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

Solution: We see that increasing the ridge parameter λ shrinks the norm of $\widehat{\theta}_\lambda$, and that even as $\lambda \to 0^+$, $\widehat{\theta}_\lambda$ picks out the least norm least squares solution.

In addition we know that adding a ridge parameter helps lower the variance of our model (in exchange for bias). High variance is caused by d_i being very close to 0, which causes $\frac{1}{d_i}$ to be large and highly variable depending on our data; small shifts in d_i cause drastic shifts in its reciprocal. With a ridge parameter we see the $\frac{1}{d_i}$ in the summation is replaced by $\frac{d_i}{d_i^2+\lambda}$, where λ is not close to 0. This new fraction becomes close to 0 when d_i is close to 0 and is much more stable; minor shifts in d_i won't cause this new value to vary drastically. This helps reduce unstable, high variance "complex" weights.

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:

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

Solution: The random variable X follows a geometric distribution with parameter b, and for all $k \in \mathbb{N}$, $\mathbb{P}(X = k) = (1-b)^{k-1}b$. By definition, its entropy (in bits) is:

$$\begin{aligned} H(X) &= -\sum_{k=1}^{\infty} (1-b)^{k-1} b \log\Big((1-b)^{k-1} b\Big) \\ &= -b\log(b) \sum_{k=1}^{\infty} (1-b)^{k-1} - b\log(1-b) \sum_{k=1}^{\infty} (k-1)(1-b)^{k-1} \\ &= -\frac{b\log(b)}{b} - b\log(1-b)\frac{1-b}{b^2} \\ &= \frac{-b\log(b) - (1-b)\log(1-b)}{b}. \end{aligned}$$

(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.

Solution: First notice that, if $b = \frac{1}{2}$, $H(X) = 2$. Now we construct a sequence of questions. Encode the answer "yes" with 1 and "no" with 0. First we ask: is $X = 1$? The answer is 1 with probability $1/2$, and 0 with the same probability. Then we ask: is $X = 2$? Conditioned the first question being answered with 0, the answer is 1 again with probability $1/2$. And similarly, at every round k, we ask: is $X = k$? This, of course, implies that the answer was 0 in all previous rounds. Therefore, the expected number of questions is exactly the entropy of $H(X)$, because requiring k questions happens with probability $\left(\frac{1}{2}\right)^k$:

$$\mathbb{E}[\text{no. questions}] = \sum_{k=1}^{\infty} k\left(\frac{1}{2}\right)^k = 2 = H(X).$$

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?

Solution: False. Example: one dimensional feature space with training points of two classes x and o arranged as xxxooooxxx. This statement would be true if the splits were allowed to form more complex boundaries, i.e. if the splits were not binary and linear.

(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.

Solution: False. Consider the XOR function, where the samples are

$$S = \{(0,0;0),(0,1;1),(1,0;1),(1,1;0)\},$$

where the first two entries in every sample are features, and the last one is the label. Then, $H(S) = 1$. The first split is done based on the first feature, which gives $S_l = \{(0,0;0),(0,1;1)\}$ and $S_r = \{(1,0;1),(1,1;0)\}$; denote the corresponding nodes as $\textbf{child}_l$ and $\textbf{child}_r$ respectively. This gives $H(S_l) = 1$ and $H(S_r) = 1$. Now we can compute the information gain of the first split:

$$IG(\textbf{root}) = H(S) - \frac{|S_l|H(S_l) + |S_r|H(S_r)}{|S_l| + |S_r|} = 0.$$

Now we further split S_l and S_r according to the second feature, which gives 4 leaves of 1 sample each. Denote the leaf samples corresponding to S_r as $L_{r,1}$ and $L_{r,2}$, and accordingly denote by $L_{l,1}$ and $L_{l,2}$ the leaves corresponding to S_l. Now we have

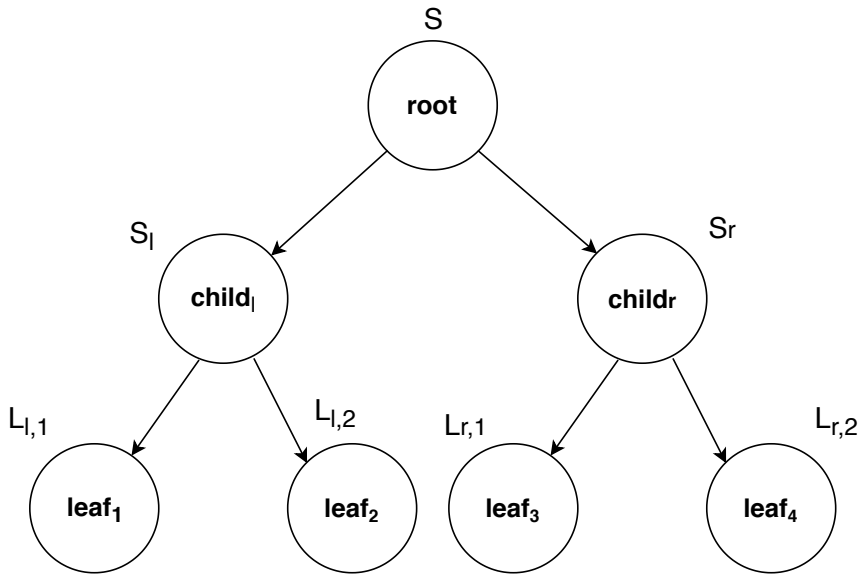
$$IG(\textbf{child}_l) = H(S_l) - \frac{1 \cdot H(L_{l,1}) + 1 \cdot H(L_{l,2})}{1+1} = 1,$$

and analogously $IG(\textbf{child}_r) = 1$. Therefore, the information gain at each of the child nodes is 1, while the information gain at the root is 0.

(c) Intuitively, how is the depth of a decision tree related to overfitting and underfitting?

Solution: If a decision tree is very deep, the model is likely to overfit. Intuitively, there are many conditions checked before making a decision, which makes the decision rule too fine-grained and sensitive to small perturbations; for example, if only one of the many conditions is not satisfied, this might result in a completely different prediction. On the other hand, if the tree is very shallow, it might underfit. In this case, the decisions are too "coarse".

(d) 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.)

Solution: Suppose that we draw n samples. Each sample has d input features plus its label, so there are 2^{d+1} distinct feature vector/label examples to choose from. For each sample, there is exactly one contradictory sample, namely the sample with the same input features but the opposite label. Thus, the probability of finding no contradiction is

$$\frac{\text{\# of sequences of non-contradictory samples}}{\text{\# of different sequences}} = \frac{2^{d+1}(2^{d+1}-1)\ldots(2^{d+1}-n+1)}{2^{n(d+1)}} = \frac{2^{d+1}!}{(2^{d+1}-n)!2^{n(d+1)}}.$$

For example, if $d = 10$, there are 2048 possible samples, and a contradiction has probability greater than 0.5 already after 54 drawn samples.