EECS 182 Deep Neural Networks
Fall 2022 Anant Sahai

Discussion 2

1. Two forms of Ridge Regression Consider the Ridge Regression estimator,

$$\underset{\mathbf{w}}{\operatorname{argmin}} \|X\mathbf{w} - y\|_2^2 + \lambda\|\mathbf{w}\|^2$$

We know this is solved by

$$\hat{\mathbf{w}} = (X^T X + \lambda I)^{-1} X^T \mathbf{y} \tag{1}$$

An alternate form of the Ridge Regression solution (often called the Kernel Ridge form) is given by

$$\hat{\mathbf{w}} = X^T (X X^T + \lambda I)^{-1} \mathbf{y}. \tag{2}$$

(a) Show that the two solutions for ridge regression are equivalent by algebraic manipulation.

(b) We know that Ridge Regression can be viewed as finding the MAP estimate when we apply a prior on the (now viewed as random parameters) $\mathbf{W}$. In particular, we can think of the prior for $\mathbf{W}$ as being $\mathcal{N}(\mathbf{0}, I)$ and view the random Y as being generated using $Y = \mathbf{x}^T\mathbf{W} + \sqrt{\lambda}N$ where the noise N is distributed iid (across training samples) as $\mathcal{N}(0,1)$. At the vector level, we have $\mathbf{Y} = X\mathbf{W} + \sqrt{\lambda}\mathbf{N}$, and then we know that when we try to maximize the log likelihood we end up minimizing

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{\lambda}\|X\mathbf{w} - y\|_2^2 + \|\mathbf{w}\|^2 = \underset{\mathbf{w}}{\operatorname{argmin}} \|X\mathbf{w} - y\|_2^2 + \lambda\|\mathbf{w}\|^2.$$

The underlying probability space is that defined by the d iid standard normals that define the $\mathbf{W}$ and the n iid standard normals that give the n different N_i on the training points. Note that the X matrix whose rows consist of the n different inputs for the n different training points are not random.

Based on what we know about joint normality, it is clear that the random Gaussian vectors $\mathbf{W}$ and $\mathbf{Y}$ are jointly normal. Use the following facts to show that the two forms of solution are identical.

- (1) is the MAP estimate for $\mathbf{W}$ given an observation $\mathbf{Y} = \mathbf{y}$.
- For jointly normal random variables, when you condition one set of variables on the values for the others, the resulting conditional distribution is still normal.
- A normal random variable has its density maximized at its mean.
- For jointly normal random vectors that are zero mean, the formula for the conditional expectation is

$$E[\mathbf{W}|\mathbf{Y} = \mathbf{y}] = \Sigma_{WY}\Sigma_{YY}^{-1}\mathbf{y} \tag{3}$$

where the Σ_{YY} is the covariance $E[\mathbf{Y}\mathbf{Y}^T]$ of $\mathbf{Y}$ and $\Sigma_{WY} = E[\mathbf{W}\mathbf{Y}^T]$ is the appropriate cross-covariance of $\mathbf{W}$ and $\mathbf{Y}$.

2. Visualizing Backpropagation Consider a simple neural network that takes a scalar real input, has 1 hidden layer with k units in it and a ReLU nonlinearity for those units, and an output linear (affine) layer.

We can algebraically write any function that it represents as

$$y = W^{(2)}(\max(\mathbf{0}, W^{(1)}x + \mathbf{b^{(1)}})) + b^{(2)}$$

Where $x, y \in \mathbb{R}$, $W^{(1)} \in \mathbb{R}^{k\times 1}$, $W^{(2)} \in \mathbb{R}^{1\times k}$, and $\mathbf{b}^{(1)} \in \mathbb{R}^{k\times 1}$, and $b^{(2)} \in \mathbb{R}$. The superscripts are indices, not exponents and the $\max$ given two vector arguments applies the max on corresponding pairs and returns a vector.

For each part, calculate the partial derivative and sketch a small representative plot of the derivative as a function of x. Make sure to clearly label any discontinuities, kinks, and slopes of segments. The subscript i refers to the i-th element of a vector.

(a) $\frac{\partial y}{\partial b^{(2)}}$

(b) $\frac{\partial y}{\partial w_i^{(2)}}$

(c) $\frac{\partial y}{\partial b_i^{(1)}}$

(d) $\frac{\partial y}{\partial w_i^{(1)}}$

3. Least Squares and the Min-norm problem from the Perspective of SVD (If time permits)

Consider the equation $X\mathbf{w} = \mathbf{y}$, where $X \in \mathbb{R}^{m\times n}$ is a non-square data matrix, w is a weight vector, and y is vector of labels corresponding to the datapoints in each row of X.

Let's say that $X = U\Sigma V^T$ is the (full) SVD of X. U and V are orthonormal square matrices, and Σ is an $m \times n$ matrix with singular values (σ_i) on the "diagonal".

For this problem, we define $\Sigma^\dagger$ an $n \times m$ matrix with the reciprocals of the singular values ($\frac{1}{\sigma_i}$) along the "diagonal".

(a) First, consider the case where $m > n$, i.e. our data matrix X has more rows than columns (tall matrix) and the system is overdetermined. How do we find the weights w that minimizes the error between $X\mathbf{w}$ and y? In other words, we want to solve $\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2$.

(b) Plug in the SVD $X = U\Sigma V^T$ and simplify. Be careful with dimensions!

(c) What happens if we left-multiply X by our least squares solution?

(d) Now, let's consider the case where $m < n$, i.e. the data matrix X has more columns than rows and the system is underdetermined. There exist infinitely many solutions for w, but we seek the minimum-norm solution, ie. we want to solve $\min \|\mathbf{w}\|^2 \text{s.t.} X\mathbf{w} = \mathbf{y}$. What is the minimum norm solution?

(e) Plug in the SVD $X = U\Sigma V^T$ and simplify. Be careful with dimensions!

(f) What happens if we right-multiply X by our min-norm solution?

