Optimization-Based Data Analysis

Recitation 8

1. Under what conditions will training error increase if you add a feature to your regression problem? How does the answer change if you are using ridge regression?

   Solution. It never increases for both.

2. Suppose you fit a linear regression model, but have scaled a feature by a factor of 10.

   (a) Under what conditions will this change the forecast $X\hat{\beta}$?

   (b) What impact will this have on ridge regression?

   Solution.

   (a) Same forecast for standard regression.

   (b) It will have the effect of reducing the penalty on the corresponding coefficient.

3. The ridge regression estimator is given by

$$\vec{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T \vec{y}.$$

   Under what conditions on X is this formula valid (i.e., does the inverse exist)?

   Solution. It always exists since $X^T X$ is positive semidefinite and λI is positive definite.

4. Let $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, $y \in \mathbb{R}^n$, $\lambda > 0$, and $M \in \mathbb{R}^{m\times p}$ has full column rank. What is the solution to

$$\arg\min_{\beta} \|X\beta - y\|_2^2 + \lambda\|M\beta\|_2^2?$$

   Solution.

$$(X^T X + \lambda M^T M)^{-1} X^T y.$$

5. Suppose you are given data $\vec{y} = X\vec{\beta} + \vec{z}$ (all variables deterministic; $\vec{\beta}, \vec{z}$ unknown) and compute the least squares estimator $\hat{\vec{\beta}}$ for $\vec{\beta}$. Assuming $\|\vec{z}\|_2 = \eta$ is fixed, and X has full column rank, what direction for $\vec{z}$ produces the largest error $\|\hat{\vec{\beta}} - \vec{\beta}\|_2$, and how much is that error?

   Solution. If $X \in \mathbb{R}^{n\times p}$ has SVD USV^T then the error $\hat{\vec{\beta}} - \vec{\beta} = US^{-1}V^T z$. This has maximum norm when the noise z points in the direction $V[:, p]$ giving a norm of η/σ_p.

6. Suppose $\vec{\mathbf{y}} = \mathbf{X}\vec{\beta} + \vec{\mathbf{z}}$ where $\mathbf{X}, \vec{\mathbf{z}}$ all have iid standard Gaussian entries. As n, the number of data points, grows, how will the error $\|\hat{\vec{\beta}} - \vec{\beta}\|_2$ decay?

Solution. Like $1/\sqrt{n}$. For large n the error concentrates around $\sqrt{p/n}$, where p is the number of features.

7. Let $\vec{\beta}_{\text{ridge}}$ denote the ridge regression estimator which minimizes

$$\arg\min_{\beta} \|X\vec{\beta} - \vec{y}\|_2^2 + \lambda\|\vec{\beta}\|_2^2.$$

Show that $\vec{\beta}_{\text{ridge}}$ is in the row space of X.

Solution. Write $\vec{\beta} = \vec{\beta}_r + \vec{\beta}_{r\perp}$ where $\vec{\beta}_r$ is the orthogonal projection of $\vec{\beta}$ onto the row space of X, and $\vec{\beta}_{r\perp}$ is the orthogonal projection of $\vec{\beta}$ onto the orthogonal complement of the row space of X. Then $X\vec{\beta} = X\vec{\beta}_r$ but $\|\vec{\beta}\|_2^2 = \|\vec{\beta}_r\|_2^2 + \|\vec{\beta}_{r\perp}\|_2^2$.

8. Suppose we have the regression problem $\vec{\mathbf{y}} = X\vec{\beta} + \vec{\mathbf{z}}$ where $\vec{\mathbf{z}}$ is iid gaussian with mean 0 and variance σ_2^2. Suppose we have a Gaussian prior on $\vec{\beta}$ with mean $\vec{\mu}$ and variance σ_1^2 (instead of mean 0). Can you guess the form of the minimization problem giving the resulting MAP estimator?

Solution. $\arg\min_\beta \|X\beta - y\|_2^2 + \lambda\|\beta - \vec{\mu}\|_2^2$

9. Compute the gradients of the following functions.

   (a) $f(\vec{x}) = \vec{w}^T\vec{x}$ where $\vec{w} = [1, 2, 3]^T$ where $f : \mathbb{R}^3 \to \mathbb{R}$.

   (b) $f(\vec{x}) = \frac{1}{2}\vec{x}^T A\vec{x} + \vec{w}^T\vec{x}$ where $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n\times n}$ and $\vec{w} \in \mathbb{R}^n$.

   (c) $f(X) = \frac{1}{2}\|X\|_F^2$ where $f : \mathbb{R}^{m\times n} \to \mathbb{R}$. When computing the gradient, treat X as a vector in $\mathbb{R}^{mn}$.

   (d) $f(X) = \det(X)$ where $f : \mathbb{R}^{m\times n} \to \mathbb{R}$. When computing the gradient, treat X as a vector in $\mathbb{R}^{mn}$.

Solution.

   (a) $\nabla f(\vec{x}) = \vec{w}$

   (b) $\nabla f(\vec{x}) = \frac{1}{2}(A + A^T)\vec{x} + \vec{w}$

   (c) $\nabla f(X) = X$

   (d) By doing an expansion along any of the rows, we see the ijth component of the gradient is the corresponding cofactor. This gives

$$\nabla f(X) = \det(X)(X^{-1})^T.$$