Recitation 10

Alex Dong

CDS, NYU

Fall 2020

Linear Regression

- Deep topic, there are entire courses on Linear Regression and friends
- Very nice to analyze mathematically. Guaranteed solutions via convexity.
- Combine with *convex* functions for regularization
- Lasso, L_1 penalty
  - Good for variable selection
- Ridge, L_2 penalty
  - The go-to baseline in most cases
- (&) In practice, don't use linear regression w/out regularization.
  - See Intro to Data Science, Machine Learning

Questions: Linear Regression Warm Up

When solving the least squares problem, the optimization problem is $min_{\beta}\|X\beta - y\|_2^2$, $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$, $\beta \in \mathbb{R}^d$. $n \geq d$

1. Explain what X, β, y represent
2. Geometrically, what are we trying to do?
3. How can we obtain the normal equation $X^T X\beta = X^T y$ from this geometric intuition?
   Hint $Im(A)^{\perp} = Ker(A^T)$
4. Under what conditions is $X^T X$ invertible? If $X^T X$ is not invertible, do the normal equations still have a solution?

Solutions 1: Linear Regression Warm Up

When solving the least squares problem, the optimization problem is $min_\beta \|X\beta - y\|_2^2$, $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$, $\beta \in \mathbb{R}^d$. $n \geq d$

Solution

1. *Explain what X, β, y represent.*
   X contains the independent variable observations on in each row, and the features in each column.
   y contains the corresponding dependent variable observations.
   β contains the coefficients that transform the independent variable into the dependent variable.

2. *Geometrically, what are we trying to do?*
   Thinking of this from the framework of linear transformations, we are trying to find a point $\hat{\beta} \in \mathbb{R}^d$, s.t $X\beta \in Im(X) \subset \mathbb{R}^n$ is closest to y.

Solutions 2: Linear Regression Warm Up

When solving the least squares problem, the optimization problem is $min_\beta \|X\beta - y\|_2^2$, $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$, $\beta \in \mathbb{R}^d$. $n \geq d$

Solution

3. *How can we obtain the normal equation $X^T X\beta = X^T y$ from this geometric intuition?*
   The point closest to y in $\mathrm{Im}(X)$ is the projection of y onto $Im(X)$. Let $P_{Im(X)}y = X\hat{\beta}$
   $$X\hat{\beta} - y \perp \mathrm{Im}(X)$$
   $$X\hat{\beta} - y \in \mathrm{Ker}(X^T) \qquad since\ Im(A)^\perp = Ker(A^T)$$
   $$X^T(X\hat{\beta} - y) = 0$$
   $$X^T X\hat{\beta} = X^T y$$
4. *Under what conditions is $X^T X$ invertible? If $X^T X$ is not invertible, do the normal equations still have a solution?*
   $X^T X$ is invertible if $rank(X) = d$. There is always a solution since $Im(X^T X) = Im(X^T)$ (use SVD, check Rec 8).

Questions: Linear Regression vs PCA

Let $(\vec{x_1}, y_1), ..., (\vec{x_n}, y_n) \in \mathbb{R}^{d+1}$ be a centered dataset. Each $\vec{x_i} \in \mathbb{R}^d$.
Let $\beta \in \mathbb{R}^d$. Let $X \in \mathbb{R}^{n \times d}$ (containing $\vec{x_1}, .., \vec{x_n}$) have full rank).
Let y be the vector containing $y_1, .., y_n$.
The OLS solution is given by $\hat{\beta} = (X^T X)^{-1} X^T y$.
We can use this to generate predictions $\hat{y} = X(X^T X)^{-1} X^T \vec{y}$.

1. Recall that $X(X^T X)^{-1} X^T$ is an orthogonal projection, which subspace is this an orthogonal projection onto?
2. Let $n = 1$; consider the subspace generated by the first principal component (from PCA) and the line generated by linear regression solution for $\vec{y} = \vec{x}\beta$. Are these the same line? If not, what is the difference?

Solutions 1: Linear Regression vs PCA

Let $(\vec{x_1}, y_1), ..., (\vec{x_n}, y_n) \in \mathbb{R}^{d+1}$ be a centered dataset.
Each $\vec{x_i} \in \mathbb{R}^d$. Let $\beta \in \mathbb{R}^n$. Let $X \in \mathbb{R}^{n \times d}$ be the design matrix.
Let y be the vector containing $y_1, .., y_n$.
The OLS solution is given by $\hat{\beta} = (X^T X)^{-1} X^T y$.
We can use this to generate predictions $\hat{y} = X(X^T X)^{-1} X^T \vec{y}$.

1. Recall that $X(X^T X)^{-1} X^T$ is an orthogonal projection, which subspace is this an orthogonal projection onto?

Solution

Using SVD, let X have SVD $X = U\Sigma V^T$, then
$X(X^T X)^{-1} X^T = UU^T$ *So* $X(X^T X)^{-1} X^T$ *is an orthogonal projection onto the columns of* U, *which span* $Im(X)$.
Note: $X \in \mathbb{R}^{n \times d}$, *and* $Im(X)$ *is a* d *dimensional subspace in* $\mathbb{R}^n$.
Question for you! How interpretable is this? (Answer... not very)

Solutions 2 : Linear Regression vs PCA

Let $(\vec{x_1}, y_1), ..., (\vec{x_n}, y_n) \in \mathbb{R}^{d+1}$ be a centered dataset.
Each $\vec{x_i} \in \mathbb{R}^d$. Let $\beta \in \mathbb{R}^n$. Let $X \in \mathbb{R}^{n \times d}$ be the design matrix.
Let y be the vector containing $y_1, .., y_n$.
The OLS solution is given by $\hat{\beta} = (X^T X)^{-1} X^T y$.
We can use this to generate predictions $\hat{y} = X(X^T X)^{-1} X^T \vec{y}$.

2. Let $n = 1$; consider the subspace generated by the first principal component (from PCA) and the line generated by linear regression solution for $\vec{y} = \vec{x}\beta$. Are these the same line? If not, what is the difference?

Solution

(Check notebook on github)
They are not the same line. PCA is an orthogonal projection that minimizes the L_2 orthogonal distance to the line, while linear regression minimizes the L_2 distance ***parallel to the y-axis*** *to the line.*

Questions: Ridge Regression

Let $X \in \mathbb{R}^{n \times d}$, $n > d$, and *not have full rank.* (X is a data matrix)
Recall that the OLS solution is $\hat{x} = (X^T X)^{-1} X^T y$.

1. Since X is not full rank, what does this say about the features?
2. What is the issue with the OLS solution?
3. The ridge regression solution is given by $(X^T X + \lambda Id_d)^{-1} X^T y$. How does this fix the issue?
4. Suppose that X has SVD $X = U \Sigma V^T$, and X has singular values $\sigma_1, ..., \sigma_d$. What are the eigenvalues of $X^T X + \lambda Id_d$?
5. How does increasing λ affect the condition number of $(X^T X + \lambda Id_d)$?

Solutions: Ridge Regression and Multicollinearity

Let $X \in \mathbb{R}^{n \times d}$, $n > d$, and *not have full rank.* (X is a data matrix)
Recall that the OLS solution is $\hat{x} = (X^T X)^{-1} X^T y$.

Solution

1. *Since X is not full rank, what does this say about the features?*
   Columns of X are not linearly independent, so some of the features can be perfectly explained by other features.
2. *What is the issue with the OLS solution?*
   Since X does not have full rank, $X^T X$ doesn't have full rank and is not invertible. So the OLS solution is not well-defined.
3. *The ridge regression solution is given by $(X^T X + \lambda Id_d)^{-1} X^T y$. How does this fix the issue?*
   Adding λId_d to $X^T X$ shifts its eigenvalues up, which makes $(X^T X + \lambda Id_d)$ invertible.

Solutions: Ridge Regression and Multicollinearity

Let $X \in \mathbb{R}^{n \times d}$, $n > d$, and *not have full rank.* (X is a data matrix)
Recall that the OLS solution is $\hat{x} = (X^T X)^{-1} X^T y$.

Solution

4. *Suppose that X has SVD $X = U\Sigma V^T$, and X has singular values $\sigma_1, ..., \sigma_d$. What are the eigenvalues of $X^T X + \lambda Id_d$?*
   Note that $X^T X = V\Sigma^T \Sigma V^T$
   Eigvals of $X^T X$: $\sigma_1^2, ..., \sigma_d^2$, (Note: X isn't full rank, so $\sigma_d = 0$)
   Eigvals of $X^T X + \lambda Id_d$: $\sigma_1^2 + \lambda, ..., \sigma_d^2 + \lambda$.
5. *How does increasing λ affect the condition number of $(X^T X + \lambda Id_d)$ vs $X^T X$?*
   Condition number of $X^T X = \frac{\sigma_1^2}{\sigma_d^2} = \infty$
   Condition number of $(X^T X + \lambda Id_d) = \frac{\sigma_1^2 + \lambda}{\sigma_d^2 + \lambda}$
   Furthermore, for $\lambda_1 > \lambda_2$, we get the relationship
   $\frac{\sigma_1^2 + \lambda_1}{\sigma_d^2 + \lambda_1} < \frac{\sigma_1^2 + \lambda_2}{\sigma_d^2 + \lambda_2}$