EECS 189 Introduction to Machine Learning
Fall 2020

DIS6

This discussion was released **Friday, October 9**.

This discussion material is on PCA, LASSO, ridge regression and their relationships. We will first start with this Jupyter notebook. We will then come back to the document and consider the relationship between PCA and ridge regression. Note that this is also part of your homework problem.

As a reminder, if you have questions, we will answer them via the queue at oh.eecs189.org. Once you complete the Jupyter notebook, please return to this worksheet.

1 Jupyter Notebook

Solution:

This jupyter notebook does not require any addtional coding.

2 Ridge regression vs. PCA

Assume we are given n training data points $(\mathbf{x}_i, y_i)$. We collect the target values into $\mathbf{y} \in \mathbb{R}^n$, and the inputs into the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ where the rows are the $d-$dimensional feature vectors $\mathbf{x}_i^\top$ corresponding to each training point. Furthermore, assume that $\frac{1}{n}\sum_{i=1}^n \mathbf{x_i} = \mathbf{0}$, $n > d$ and $\mathbf{X}$ has rank d.

In this problem we want to compare two procedures: The first is ridge regression with hyperparameter λ, while the second is applying ordinary least squares after using PCA to reduce the feature dimension from d to k (we give this latter approach the short-hand name k-PCA-OLS where k is the hyperparameter).

Notation: The singular value decomposition of $\mathbf{X}$ reads $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{n\times n}$, $\mathbf{\Sigma} \in \mathbb{R}^{n\times d}$ and $\mathbf{V} \in \mathbb{R}^{d\times d}$. We denote by $\mathbf{u}_i$ the n-dimensional column vectors of $\mathbf{U}$ and by $\mathbf{v}_i$ the $d-$dimensional column vectors of $\mathbf{V}$. Furthermore the diagonal entries $\sigma_i = \Sigma_{i,i}$ of $\mathbf{\Sigma}$ satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > 0$. For notational convenience, assume that $\sigma_i = 0$ for $i > d$.

(a) It turns out that the ridge regression optimizer (with $\lambda > 0$) in the $\mathbf{V}$-transformed coordinates

$$\widehat{\mathbf{w}}_{\text{ridge}} = \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{V}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

has the following expression:

$$\widehat{\mathbf{w}}_{\text{ridge}} = \text{diag}(\frac{\sigma_i}{\lambda + \sigma_i^2})\mathbf{U}^\top\mathbf{y}. \tag{1}$$

DIS6, ©UCB EECS 189, Fall 2020. All Rights Reserved. This may not be publicly shared without explicit permission.

Use $\widehat{y}_{test} = \mathbf{x}_{test}^\top \mathbf{V}\widehat{\mathbf{w}}_{\text{ridge}}$ to denote the resulting prediction for a hypothetical $\mathbf{x}_{test}$. Using (1) and the appropriate scalar $\{\beta_i\}$, this can be written as:

$$\widehat{y}_{test} = \mathbf{x}_{test}^\top \sum_{i=1}^{d} \mathbf{v}_i \beta_i \mathbf{u}_i^\top \mathbf{y}. \tag{2}$$

What are the $\beta_i \in \mathbb{R}$ for this to correspond to (1) from ridge regression?

Solution:

The resulting prediction for ridge reads

$$\begin{aligned}\hat{\mathbf{y}}_{\text{ridge}} &= \mathbf{x}^\top \mathbf{V}\,\text{diag}\left(\frac{\sigma_i}{\lambda + \sigma_i^2}\right)\mathbf{U}^\top \mathbf{y} \\ &= \mathbf{x}^\top \sum_{i=1}^{d} \frac{\sigma_i}{\lambda + \sigma_i^2} \mathbf{v}_i \mathbf{u}_i^\top \mathbf{y}\end{aligned}$$

Therefore we have $\beta_i = \frac{\sigma_i}{\lambda + \sigma_i^2}$ for $i = 1, \ldots, d$.

(b) Suppose that we do k-PCA-OLS — i.e. ordinary least squares on the reduced k-dimensional feature space obtained by projecting the raw feature vectors onto the $k < d$ principal components of the covariance matrix $\mathbf{X}^\top \mathbf{X}$. Use $\widehat{y}_{test}$ to denote the resulting prediction for a hypothetical $\mathbf{x}_{test}$,

It turns out that the learned k-PCA-OLS predictor can be written as:

$$\widehat{y}_{test} = \mathbf{x}_{test}^\top \sum_{i=1}^{d} \mathbf{v}_i \beta_i \mathbf{u}_i^\top \mathbf{y}. \tag{3}$$

Give the $\beta_i \in \mathbb{R}$ coefficients for k-PCA-OLS. Show work.

Hint 1: some of these β_i will be zero. Also, if you want to use the compact form of the SVD, feel free to do so if that speeds up your derivation.

Hint 2: some inspiration may be possible by looking at the next part for an implicit clue as to what the answer might be.

Solution: The OLS on the k-PCA-reduced features reads

$$\min_{\mathbf{w}} \|\mathbf{X}\mathbf{V}_k \mathbf{w} - \mathbf{y}\|_2^2$$

where the columns of $\mathbf{V}_k$ consist of the first k eigenvectors of $\mathbf{X}$.

In the following we use the compact form SVD, that is note that one can write

$$\begin{aligned}\mathbf{X} &= \mathbf{U}\mathbf{\Sigma}\mathbf{V} \\ &= \mathbf{U}_d \mathbf{\Sigma}_d \mathbf{V}\end{aligned}$$

where $\mathbf{\Sigma}_d = \text{diag}(\sigma_i)$ for $i = 1, \ldots, d$ and $\mathbf{U}_d$ are the first d columns of $\mathbf{U}$. In general we use the notation $\mathbf{\Sigma}_k = \text{diag}(\sigma_i)$ for $i = 1, \ldots, k$.

DIS6, ©UCB EECS 189, Fall 2020. All Rights Reserved. This may not be publicly shared without explicit permission.

Apply OLS on the new matrix $\mathbf{X}\mathbf{V}_k$ to obtain

$$\begin{aligned}\widehat{\mathbf{w}}_{\mathrm{PCA}} &= [(\mathbf{X}\mathbf{V}_k)^\top(\mathbf{X}\mathbf{V}_k)]^{-1}(\mathbf{X}\mathbf{V}_k)^\top\mathbf{y} \\ &= [\mathbf{V}_k^\top\mathbf{V}\mathbf{\Sigma}_d^2\mathbf{V}^\top\mathbf{V}_k]^{-1}\mathbf{V}_k^\top\mathbf{X}^\top\mathbf{y} \\ &= \mathbf{\Sigma}_k^{-1}\mathbf{U}_k^\top\mathbf{y} = \widetilde{\mathbf{\Sigma}}_k^{-1}\mathbf{U}^\top\mathbf{y}\end{aligned}$$

where $\widetilde{\mathbf{\Sigma}}_k = \begin{pmatrix}\mathbf{\Sigma}_k & 0\end{pmatrix}$

The resulting prediction for PCA reads (note that you need to project it first!)

$$\begin{aligned}\widehat{\mathbf{y}}_{\mathrm{PCA}} &= \mathbf{x}^\top\mathbf{V}_k\widehat{\mathbf{w}}_{\mathrm{PCA}} \\ &= \mathbf{x}^\top\mathbf{V}_k\mathbf{\Sigma}_k^{-1}\mathbf{U}_k^\top\mathbf{y} \\ &= \mathbf{x}^\top\sum_{i=1}^{k}\frac{1}{\sigma_i}\mathbf{v}_i\mathbf{u}_i^\top\mathbf{y}\end{aligned}$$

and hence $\beta_i = \frac{1}{\sigma_i}$ if $i \leq k$ and $\beta_i = 0$ for $i = k+1, \ldots, d$.

(c) Compare $\widehat{\mathbf{y}}_{\mathrm{PCA}}$ with $\hat{\mathbf{y}}_{\mathrm{ridge}}$ (at different λ), how do you find their relationship? **Solution:**

(a) If $\lambda = 0$, ridge regression degenerates to ordinary least squares.

(b) If $\lambda > 0$, the larger the singular value σ_i, the less it will be penalized in ridge regression.

(c) In contrast for k-PCA-OLS (PCA regression), large singular values are kept intact, while samll ones (after certain number k) are completely removed. This would correspond to $\lambda = 0$ for the first k components and $\lambda = \infty$ for the rest.

(d) This means that the regression can be considered as a "smooth version" of PCA regression.

DIS6, ©UCB EECS 189, Fall 2020. All Rights Reserved. This may not be publicly shared without explicit permission.

Contributors:

- Anant Sahai
- Fanny Yang
- Peter Wang
- Philipp Moritz