CS 189/289A Introduction to Machine Learning
Fall 2022 Jennifer Listgarten, Jitendra Malik **DIS6**

1 Derivation of 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 the data has been centered such that $\frac{1}{n}\sum_{i=1}^n \mathbf{x_i} = \mathbf{0}$, $n > d$ and $\mathbf{X}$ has rank d. The covariance matrix is given by

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

When $\bar{\mathbf{x}} = 0$ (i.e., we have subtracted the mean in our samples), we obtain $\Sigma = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$. We will assume this to be the case for this problem.

(a) Maximum Projected Variance: We would like the vector $\mathbf{w}$ such that projecting your data onto $\mathbf{w}$ will retain the maximum amount of information, i.e., variance. We can formulate the optimization problem as

$$\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^\top\mathbf{w}\right)^2 = \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\,. \quad (1)$$

Show that the maximizer for this problem is equal to the eigenvector $\mathbf{v}_1$ that corresponds to the largest eigenvalue λ_1 of Σ. Also show that the optimal value of this problem is equal to λ_1.

Hint: Use the spectral decomposition of Σ and consider reformulating the optimization problem using a new variable.

Solution:

We start by invoking the spectral decomposition of $\Sigma = \mathbf{V}\Lambda\mathbf{V}^\top$, which is a symmetric positive semi-definite matrix.

$$\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} = \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top\mathbf{V}\Lambda\mathbf{V}^\top\mathbf{w} = \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} (\mathbf{V}^\top\mathbf{w})^\top\Lambda\mathbf{V}^\top\mathbf{w} \quad (2)$$

Define a new variable $\mathbf{z} = \mathbf{V}^\top\mathbf{w}$, and maximize over this variable. Note that because $\mathbf{V}$ is invertible, there is a one to one mapping between $\mathbf{w}$ and $\mathbf{z}$. Also note that the constraint is the same because the length of the vector $\mathbf{w}$ does not change when multiplied by an orthogonal matrix.

$$\max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \mathbf{z}^\top\Lambda\mathbf{z} = \max_{\mathbf{z}:\|\mathbf{z}\|_2=1} \sum_{i=1}^{d}\lambda_i z_i^2$$

From this new formulation, we can see that we can maximize this by "throwing all of our eggs into one basket"; that is, setting $z_i^* = 1$ if i is the index of the largest eigenvalue, and $z_i^* = 0$ otherwise. Note that, under our constraint that the norm of z must be 1, this maximizes our value: if we were to reduce the value of z_i that corresponds to the largest eigenvalue and assign that value to a different z_j with a smaller or equal eigenvalue, we would get a value that is strictly less than or equal to setting z_i to 1. In other words, $\mathbf{z}$ is a one hot vector. Thus,

$$\mathbf{z}^* = \mathbf{V}^\top \mathbf{w}^* \implies \mathbf{w}^* = \mathbf{V}\mathbf{z}^* = \mathbf{v}_1$$

where $\mathbf{v}_1$ is the principal eigenvector and corresponds to λ_1. Plugging this into the objective function, we see that the optimal value is λ_1.

(b) Let us call the solution of the above part $\mathbf{w}_1$. Next, we will use a *greedy procedure* to find the ith component of PCA by doing the following optimization

$$\begin{aligned} \text{maximize} \quad & \mathbf{w}_i^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}_i \\ \text{subject to} \quad & \mathbf{w}_i^\top \mathbf{w}_i = 1 \\ & \mathbf{w}_i^\top \mathbf{w}_j = 0 \quad \forall j < i, \end{aligned} \tag{3}$$

where $\mathbf{w}_j, j < i$ are defined recursively using the same maximization procedure above. Show, using your work in the previous part, that the maximizer for this problem is equal to the eigenvector $\mathbf{v}_i$ that corresponds to the ith eigenvalue λ_i of Σ. Also show that optimal value of this problem is equal to λ_i.

Solution: We can use the same strategy as from the previous part to write the optimization problem as

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{i=1}^d \lambda_i z_i^2 \\ \text{subject to} \quad & \|\mathbf{z}\|_2 = 1 \\ & \mathbf{z}_j = 0 \quad \forall j < i, \end{aligned} \tag{4}$$

We see that we can maximize this by throwing all of our eggs into one basket, as explained in the previous part, and setting $z_k^* = 1$ if k is the index of the ith largest eigenvalue and others to 0. Plugging this into the objective function, we see that the optimal value is λ_i.

2 Ridge regression vs. PCA

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) Consider running ridge regression with $\lambda > 0$ in the $\mathbf{V}$-transformed coordinates, i.e.,

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

Note that this does not correspond to any dimensionality reduction, just a change of variables. It turns out that the solution in this case can be written as:

$$\widehat{\mathbf{w}}_{\text{ridge}} = \left[\operatorname{diag}\left(\frac{\sigma_1}{\lambda + \sigma_1^2}, \ldots, \frac{\sigma_d}{\lambda + \sigma_d^2}\right) \; 0\right] \mathbf{U}^\top \mathbf{y}. \tag{5}$$

The matrix notation above refers to a diagonal matrix, where the first d dimensions have diagonal entries $\frac{\sigma_i}{\lambda+\sigma_i^2}$ for some dimension $i \leq d$, and the rest of the dimensions are 0 for $j > d$. 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 (5) and the appropriate scalar $\{\beta_i\}$ (find the value for this), show that this prediction can be written as:

$$\widehat{y}_{test} = \mathbf{x}_{test}^\top \sum_{i=1}^{d} \mathbf{v}_i \beta_i \mathbf{u}_i^\top \mathbf{y}. \tag{6}$$

Solution:

The resulting prediction for ridge reads

$$\begin{aligned}
\hat{\mathbf{y}}_{\text{ridge}} &= \mathbf{x}^\top \mathbf{V} \left[\operatorname{diag}\left(\frac{\sigma_1}{\lambda + \sigma_1^2}, \ldots, \frac{\sigma_d}{\lambda + \sigma_d^2}\right) \; 0\right] \mathbf{U}^\top \mathbf{y} \\
&= \mathbf{x}^\top \sum_{i=1}^{d} \frac{\sigma_i}{\lambda + \sigma_i^2} \mathbf{v}_i \mathbf{u}_i^\top \mathbf{y}
\end{aligned}$$

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 Σ. 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 also be written as:

$$\widehat{y}_{test} = \mathbf{x}_{test}^\top \sum_{i=1}^{d} \mathbf{v}_i \beta_i \mathbf{u}_i^\top \mathbf{y}. \tag{7}$$

What are the $\beta_i \in \mathbb{R}$ coefficients in this case?

Hint: Some of these β_i will be zero.

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 $\mathbf{V}_k$ denotes the first k columns of $\mathbf{V}$.

In the following, we use the compact form SVD, that is:

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

where $\mathbf{\Sigma}_d = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ and $\mathbf{U}_d$ are the first d columns of $\mathbf{U}$. In general we use the notation $\mathbf{\Sigma}_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$.

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

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

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 regularization values λ, how does the relationship between the two vary?

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 small 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 ridge regression can be thought of as a "smooth version" of PCA regression.