10-601A Fall 2015
HW 3 Written Solutions

Cole Gulino

November 18, 2015

2. Kernel PCA

In this problem, we will derive the kernelized form of PCA, i.e., find a low-dimensional linear projection of a high-dimensional (usually nonlinear) transformation of the original features. We will assume for now that the data is pre-centered, i.e., the mean of each feature is 0. Recall that the principal components are eigenvectors of $\mathbf{X}\mathbf{X}^T$, where columns of $\mathbf{X}$ are the n data points, i.e., $\mathbf{X} = [x_1 x_2 \ldots x_n]$ where each $x_i \in \mathbb{R}^P$.

Show that any principle component $\mathbf{w}$ can be written as a linear combination of the data points $\{x_i\}_{i=1}^n$, i.e. $\mathbf{w} = \mathbf{X}\mathbf{a}$

We can consider a rank q linear model for representing the observations of $\mathbf{X}$:
$f(\lambda) = \mu + \mathbf{V}_q \lambda$

Where

- μ is a location vector in $\mathbb{R}^p$
- $\mathbf{V}_q$ is a p x p matrix q x q orthogonal unit vectors as columns
- λ is a q vector of parameters

This is the parametric representation of an affine hyperplane of rank q

Fitting such a model to the data by least squares amounts to minimizing the reconstruction error:

$$\min_{\mu, \{\lambda_i\}, \mathbf{V}_q} \sum_{i=1}^{N} ||x_i - \mu - \mathbf{V}_q \lambda_i||^2$$

We can say that $\hat{\mu}$ and $\hat{\lambda}_i$ equal:

$$\hat{\mu} = \bar{x} = \text{mean of } x$$

$$\hat{\lambda}_i = \mathbf{V}_q^T (x_i - \bar{x})$$

So then the reconstruction error becomes:

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||\,(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})\,||^2$$

The assumption above that the mean of each feature is zero, so $\bar{x}$ = 0

$$\underset{\mathbf{V}_q}{min} \sum_{i=1}^{N} ||x_i - \mathbf{V}_q \mathbf{V}_q^T x_i||^2$$

We can say that the p x p matrix: $\mathbf{H}_q = \mathbf{V}_q \mathbf{V}_q^T$ maps each point onto its rank-q reconstruction

This is the solution that we are trying to find.

So we can construct the singular value decomposition of $\mathbf{X}$
$\mathbf{X} = \mathbf{UDV}^T$

$\mathbf{U}$ is a matrix containing the eigenvectors of $\mathbf{XX}^T$

From the information above, we see that $\mathbf{w} = \mathbf{U}$

So then we can say that: $\mathbf{X} = \mathbf{wDV}^T$

Solving for w:
$\mathbf{X}\left(\mathbf{DV}^T\right)^{-1} = \mathbf{wDV}^T\left(\mathbf{DV}^T\right)^{-1}$
$\mathbf{w} = \mathbf{X}\left(\mathbf{DV}^T\right)^{-1}$

Now say: $\mathbf{w}_{pxn} = \mathbf{X}_{pxn}\left(\mathbf{D}_{nxn}\mathbf{V}_{nxn}^T\right)^{-1}$
So we see that the principal component is just a linear combination of the data points $\mathbf{X}$.
We can see that if we only look at one of the principal components:
$\mathbf{w}_{i(px1)} = \mathbf{X}_{(pxn)}\text{ith column}\left[\left(\mathbf{D}_{(nxn)}\mathbf{V}_{(nxn)}^T\right)^{-1}\right]_{(nx1)}$
So we can see that the ith principal component corresponds to the linear combination of the data points and the ith column of the matrix $\left(\mathbf{DV}^T\right)^{-1}$

So we can say that for any principal component:
$\mathbf{w} = \mathbf{Xa}$
where: $\mathbf{a} = \text{ith column}\left[\left(\mathbf{DV}^T\right)^{-1}\right]$

Using the above, write the objective of maximizing variance between projected data points (data points projected on $\mathbf{w}$) as an optimization over $\mathbf{a}$ that only involves $\mathbf{a}$ and the n × n Gram matrix $\mathbf{G}$ of all pairwise inner products between the n data points.

The Gram matrix $\mathbf{G}$ for finite dimensional real vectors with a normal Euclidean dot product [1] is simply:
$\mathbf{G} = \mathbf{V}^T\mathbf{V}$

In our case, the data satisfies all of the points above, so the Gram matrix for us will be:
$\mathbf{G} = \mathbf{X}^T\mathbf{X}$

The objective of maximizing the variance between projected data points takes the form:

$$\sigma^2 = \frac{1}{N}\sum_i (\mathbf{x}_i - \mathbf{w})^2$$

$$\sigma^2 = \mathbf{w}^T \left(\mathbf{X}\mathbf{X}^T\right)\mathbf{w}$$

So we want to maximize: $\sigma^2 = \mathbf{w}^T \left(\mathbf{X}\mathbf{X}^T\right)\mathbf{w}$
Such that $\mathbf{w}^T\mathbf{w} = 1$

$$\sigma^2 = (\mathbf{X}\mathbf{a})^T \left(\mathbf{X}\mathbf{X}^T\right)(\mathbf{X}\mathbf{a})$$

$$\sigma^2 = \mathbf{a}^T\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)(\mathbf{X}\mathbf{a})$$

$$\sigma^2 = \mathbf{a}^T \left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{a}$$

$\sigma^2 = \mathbf{a}^T\mathbf{G}^2\mathbf{a}$ is what we are maximizing over.

$$\hat{\mathbf{a}} = \max_{\mathbf{a}} \left(\mathbf{a}^T\mathbf{G}^2\mathbf{a}\right)$$

The solution $\mathbf{a}$ is no longer the principal component, but argue that it is the vector of projections of data points onto the principal component, normalized by the variance of data points in that direction.

So from what we have proved in the first part, it is obvious that the $\mathbf{a}$ vector corresponds to something that projects the data points onto the principal component.

This comes from the proof we just gave that: $\mathbf{w} = \mathbf{X}\mathbf{a}$

We can see that the $\mathbf{a}$ vector must in some way project the data points onto the principal components or the equation: $\mathbf{w} = \mathbf{X}\mathbf{a}$ would not hold.

If we look again at the singular value decomposition:
$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$

We remember that
$\mathbf{a} = \left(\mathbf{D}\mathbf{V}^T\right)^{-1}$

For any given column of $\left(\mathbf{D}\mathbf{V}^T\right)^{-1}$, $\mathbf{a}$ corresponds to one of the principal components $\mathbf{w}$. What $\mathbf{a}$ then gives is an eigenvalue λ, which corresponds to the magnitude of the principal component axis.

If you look at the SVD and remember that $\mathbf{w} = \mathbf{U}$
$\mathbf{U}^{-1}\mathbf{X} = \mathbf{D}\mathbf{V}^T$
$\left(\mathbf{D}\mathbf{V}^T\right)^{-1} = \left(\mathbf{U}^{-1}\mathbf{X}\right)^{-1}$
$\mathbf{a} = \left(\mathbf{U}^{-1}\mathbf{X}\right)^{-1}$

So here you can see that a is the projection of the principal component $\mathbf{w} = \mathbf{U}$.
And it is normalized by the eigenvalues of the eigenvectors that correspond to the principal component being looked at.

If the data was not centered, argue that you can modify the objective by replacing the Gram matrix $\mathbf{G}$ with $\mathbf{G} - \frac{1}{n}\mathbf{1}^T\mathbf{G} - \frac{1}{n}\mathbf{G}\mathbf{1} + \frac{1}{n^2}\mathbf{1}^T\mathbf{G}\mathbf{1}$ where 1 is the all ones n × n matrix, and still write the solution in terms of the Gram matrix only.

If the matrix is not centered, we need to first center it, so we can say:

$$\mathbf{X}' = \mathbf{X} - \bar{\mathbf{X}}$$

Where:

$$\bar{\mathbf{X}} = \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_n \end{bmatrix}$$

And: μ_{px1} is a px1 matrix that is the average value of each of the p features of each of the n data points

$$\mu = \frac{1}{n}\left[\dots\right]_{px1}$$

A way to generalize it further is to say that:

$$\bar{\mathbf{X}} = \frac{1}{n}\mathbf{X}\mathbf{1}$$

So then:

$$\mathbf{X}' = \mathbf{X} - \frac{1}{n}\mathbf{X}\mathbf{1}$$

So now if I want to bring this into the Gram Matrix $\mathbf{G}$

I can say that:

$$\mathbf{X}'^T\mathbf{X}' = \left(\mathbf{X} - \frac{1}{n}\mathbf{X}\mathbf{1}\right)^T\left(\mathbf{X} - \frac{1}{n}\mathbf{X}\mathbf{1}\right)$$

$$\left(\mathbf{X} - \frac{1}{n}\mathbf{X}\mathbf{1}\right)^T\left(\mathbf{X} - \frac{1}{n}\mathbf{X}\mathbf{1}\right) = \mathbf{X}^T\mathbf{X} - \frac{1}{n}\mathbf{1}^T\mathbf{X}^T\mathbf{X} - \frac{1}{n}\mathbf{X}^T\mathbf{X}\mathbf{1} + \frac{1}{n^2}\mathbf{1}^T\mathbf{X}^T\mathbf{X}\mathbf{1}$$

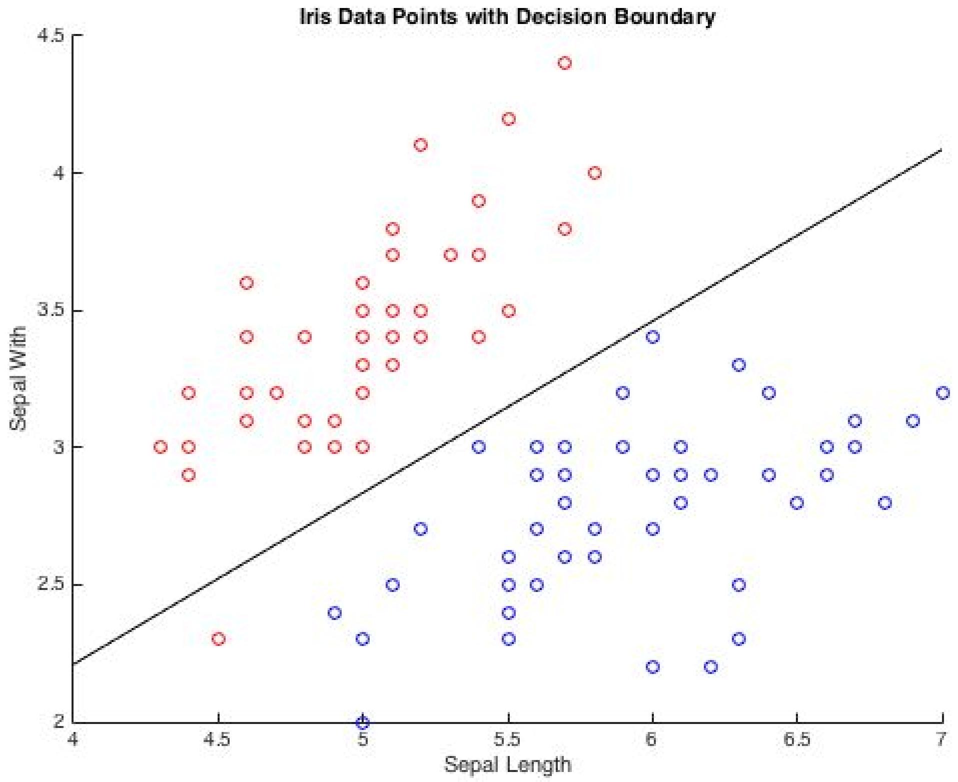
Now you can replace: $\mathbf{G} = \mathbf{X}^T\mathbf{X}$

$$\mathbf{G}' = \mathbf{G} - \frac{1}{n}\mathbf{1}^T\mathbf{G} - \frac{1}{n}\mathbf{G}\mathbf{1} + \frac{1}{n^2}\mathbf{1}^T\mathbf{G}\mathbf{1}$$

So this proves that if we make this replacement for $\mathbf{G}$, we can still write the solution in terms of the Gram matrix only, even if the data is not centered.

Problem 3 Part 3

Shown below is the visualization of the SVM decision boundary



Resources

[1] https://en.wikipedia.org/wiki/Gramian_matrix
[2] http://arxiv.org/pdf/1404.1100.pdf
[3] The Elements of Statistical Learning: Data Mining, Inference, and Prediction
[4] http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf