Tutorial 13 : Misc. Topics

1 Kernels + PCA

(a) Consider a matrix $X \in \mathbb{R}^{d\times n}$ whose columns are centered samples $x_i \in \mathbb{R}^d$ ($\hat{\mu} = 0$). Define the covariance matrix $C = XX^T$ and the inner product matrix $K = X^T X$. Prove that if a vector a is an eigenvector of K with eigenvalue λ then the vector $v \equiv Xa$ is an eigenvector of C with the same eigenvalue. Assume that $a, v \neq 0$.

(b) Assuming that a is a normalized eigenvector of K, find the norm v.

(c) In class we saw that the projection onto the mth principle component is given by $v_m^T X$ where v is an eigenvector of C which corresponds to the mth eigenvalue. Find the projection of $X = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$ on the first principle component when using a transofrmation given by the kernel $K(x_1, x_2) = (1 + x_1^T x_2)^2$.

Solution

(a) Assume $Ka = \lambda a$:

$$\begin{aligned} &Ka = \lambda a \\ \Rightarrow\ &Ka = X^T X a = X^T v = \lambda a \\ \Rightarrow\ &Cv = XX^T v = \lambda X a = \lambda v. \end{aligned}$$

(b) The norm of v is given by

$$||v||_2^2 = v^T v = (Xa)^T (Xa) = a^T X^T X a = a^T K a = \lambda a^T a = \lambda ||a||_2^2 = \lambda.$$

(c) Let v_1 be the normalized eigenvector of $\tilde{C}$ which corresponds to the first principle component in the feature space. From (b) we have that $v_1 = \frac{1}{\sqrt{\lambda}} \tilde{X} a_1$ where a_1 is the eigenvector of $\tilde{K}$ which correspond to the largest eigenvalue. Thus, we get

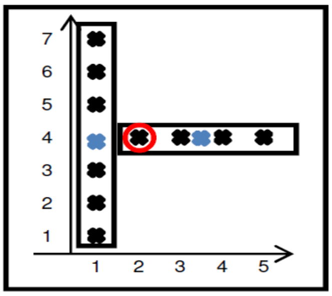
$$v_1^T \tilde{X} = \frac{1}{\sqrt{\lambda}} a_1^T \tilde{X}^T \tilde{X} = \frac{1}{\sqrt{\lambda}} a_1^T \tilde{K} = \frac{1}{\sqrt{\lambda}} (Ka)^T = \sqrt{\lambda} a_1^T.$$

In addition, we have that $\tilde{K}_{ij} = (1 + x_i^T x_j)^2 \ \Rightarrow \ \begin{bmatrix} 9 & 1 \\ 1 & 9 \end{bmatrix}$. The eigenvalues are $\lambda_{1,2} = 10, 8$ and the eigenvector related to $\lambda = 10$ is $a_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Overall, we get the projections onto the first principle component are

$$\frac{1}{\sqrt{10}} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} 9 & 1 \\ 1 & 9 \end{bmatrix} = \sqrt{5} \begin{bmatrix} 1 \\ 1 \end{bmatrix}^T.$$

2 K-Means

Consider the following set of points: For $K = 2$, is there any initialization for which K-Means will converge to two cluster marked by the black rectangles? Explain.



Solution

There is no such initialization. Let assume by contradiction there is, then, the points marked in blue will be the centroids. In this case, the point marked by a red circle will be classified to the class on the left, hence, the algorithm will not converge to this solution.

3 Kernels

Consider two kernels k_1, $k_2 : X \times X \to \mathbb{R}$. It is given that the classification problem is linearly separable for k_1 but not for k_2. Determine if the following functions are kernel functions and for which of the kernels the problem is linearly separable.

(a) $k_3(x, z) = k_1(x, z) + k_2(x, z)$.

(b) $k_4(x, z) = k_1^2(x, z)$.

Solution

(a) First we show that the function is a kernel function

- Symmetry - $k_3(x, z) = k_1(x, z) + k_2(x, z) = k_1(z, x) + k_2(z, x) = k_3(z, x)$.
- PSD - $\sum_{i,j} k_3(x_i, x_j)c_i c_j = \sum_{i,j} k_1(x_i, x_j)c_i c_j + \sum_{i,j} k_2(x_i, x_j)c_i c_j \geq 0 \; \forall c_i, c_j \in \mathbb{R}$.

The problem is linearly separable for k_3. Define $\phi_3(x) = [\phi_1(x), \phi_2(x)]$. Then,

$$\phi_3(x)^T \phi_3(z) = [\phi_1(x), \phi_2(x)]^T [\phi_1(z), \phi_2(z)] = \phi_1(x)^T \phi_1(z) + \phi_2(x)^T \phi_2(z) = k_1(x, z) + k_2(x, z).$$

Therefore, we can obtain linear separation by zeroing the coefficients of $\phi_2(x)$.

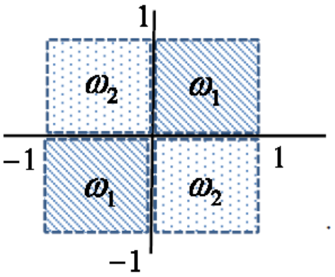
(b) $k_4(x, z)$ is a kernel function -

- Symmetry - $k_4(x, z) = k_1^2(x, z) = k_1^2(z, x) = k_4(z, x)$.
- PSD - $\sum_{i,j} k_4(x_i, x_j)c_i c_j = \sum_{i,j} k_1^2(x_i, x_j)c_i c_j \geq 0 \; \forall c_i, c_j \in \mathbb{R}$.

The problem is not linearly separable for k_4. For example, assume that $d = 1$ and $\phi_1(x) \in \{-1, 1\}$ the problem is linear separable. However, for $\phi_4(x) = x^2 = 1$ it is clear the there in

4 Bayes Decision Rule

Consider a binary classification problem with input space $X = \mathbb{R}^2$ and output space $\Omega = \{\omega_1.\omega_2\}$. It is given that $p(\omega_1) = p(\omega_2) = 0.5$ and $p(x|\omega_1)$ is a uniform distribution on quarters I and III, and $p(x|\omega_2)$ is a uniform distribution on quarters II and IV: For each of the following classifiers,



determine what is the generalization error (i.e., what is the error of an infinite set of examples drawn from the distributions above).

(a) Optimal Bayes classifier (based on the true distributions).

(b) Naive Bayes classifier which assumes a normal distribution in each one of the axes for each class. The parameters (mean and variance) are estimated using a maximum likelihood estimator from an infinite set of examples.

(c) K-NN classifier with $K = 1$ and euclidean metric which is based on two examples: $x_1 = (0.5, 0.5) \in \omega_1$ and $x_2 = (-0.5, 0.5) \in \omega_2$.

(d) K-NN classifier with $K = 1$ and euclidean metric which is based on two examples: $x_1 = (1, 1) \in \omega_1$ and $x_2 = (0, 0) \in \omega_2$.

Solution

(a) The optimal Bayes' classifier is given by

$$\hat{\omega}(x) = \arg\max_{\omega} p(x|\omega)p(\omega).$$

Since there is not overlap between the distributions of the two classes, the optimal Bayes classifier will always classify correctly, i.e., the generalization error is 0.

(b) From symmetry we get the maximum likelihood estimator of the mean is zero for each axes and each class. In addition, the maximum likelihood estimator of the variance will be the same for each axes and each class. Thus, we get two identical Gaussian distributions for both classes, hence, the classifier will classify to ω_1 or ω_2 arbitrary, regardless to the value of the sample, and the generalization error will be 50%.

(c) In this case, the decision boundary will be the y axis. This classifier will be wrong for all examples from quarters III and IV, i.e., the generalization error is 50%.

(d) In this case, the decision boundary will be the lines $y = 1 - x$ where the area above it will be classified to ω_1 and area beneath it to ω_2. The classifier will be wrong on examples taken from quarter III and from half of quarter I. Overall the generalization error will be $\frac{1.5}{4} = \frac{3}{8} = 37.5\%$.