Sharif University of Technology
Computer Engineering Department
Ali Sharifi-Zarchi

Homework Assignment 1
Machine Learning for Bioinformatics,
Spring 2023

1401-12-11

Ali Shafiei, Fakhredin Abdi

1 Linear Regression

Given n training data with m features, let the target value vector be $y = [y^{(0)}, ..., y^{(n)}] \in \mathbb{R}^n$ and data samples be $X = [x^{(0)}; ...; x^{(n)}] \in \mathbb{R}^{n\times m}$. In this context, x_j denotes the jth column of this matrix.

1.1

Show that if we train the regressor on just one of the features (from m features), we then have:

$$w_j = \frac{x_j^T y}{x_j^T x_j}$$

1.2

Suppose that the columns of matrix X are orthogonal. Prove that the optimal parameters from training the regressor on all features are the same as the optimal parameters resulting from training on each feature independently.

2 PCA

Suppose we do PCA, projecting each x_i into $z_i = V_{1:k}^T x_i$ where $V_{1:k} = [v_1, ..., v_k]$, i.e., the first k principal components. We can reconstruct x_i from z_i as $\hat{x_i} = V_{1:k} z_i$.

2.1

Show that $\|\hat{x_i} - \hat{x_j}\| = \|z_i - z_j\|$.

2.2

Show that the error in the reconstruction equals:

$$\Sigma_{i=1}^n \|x_i - \hat{x_i}\|_2^2 = (n-1)\Sigma_{j=k+1}^p \lambda_j$$

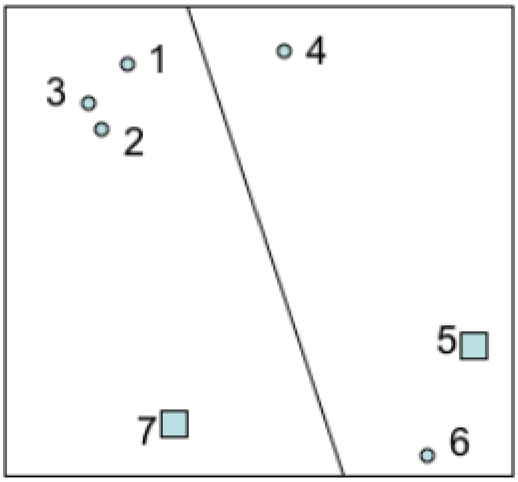
where $\lambda_{k+1}, ..., \lambda_p$ are the p-k smallest eigenvalues. Thus, the more principal components we use for the reconstruction, the more accurate it is. Further, using the top k principal components is optimal in the sense of the least reconstruction error.

3 K-means

Perform K-means on the dataset given below. Circles are data points and there are two initial cluster centers, at data points shown with squares.

3.1

Draw the cluster centers (as squares) and the decision boundaries that define each cluster. Use as many of the pictures as you need for convergence.



3.2

What is the advantage of hierarchical clustering and K-means over each other (one item for each)?

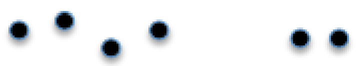
4 Gaussian Mixture Model (GMM)

Suppose that our GMM is a mixture of two Gaussians:

$$p(x) = \pi_0 N(\mu_0, \sigma_0 I) + (1 - \pi_0) N(\mu_1, \sigma_1 I)$$

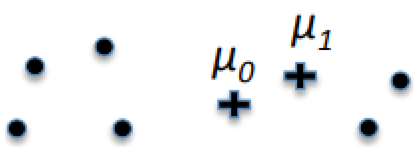
4.1

Consider the set of training data below, and two clustering algorithms: K-Means, and GMM using EM (Expectation Maximization). Will these algorithms produce the same cluster centers?



4.2

Consider applying EM to train a Gaussian Mixture Model (GMM) to cluster the data below into two clusters. The '+' points indicate the current means μ_0 and μ_1 of the two Gaussian mixture components after the k-th iteration of EM.



4.2.1

In which direction μ_0 and μ_1 will move during the next M-step?

4.2.2

Will the marginal likelihood of data, increase or decrease on the next EM iteration?