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:

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

Solution:

In this particular case, it looks like our data matrix is equal to x_j . So according to the linear regression relation we have equality below:

$$w_j = (x_j^T x_j)^{-1} x_j y = \frac{x_j^T y}{x_j^T x_j}$$

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.

Solution:

Note that the columns are orthogonal therefore their internal multiplication is zero. As a result, the value of $X^T X$ will be diagonal. More precisely:

$$X^T X = diag(x_1^T x_1, ..., x_m^T x_m) \rightarrow (X^T X)^{-1} = diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1})$$

now we have :

$$w = (X^T X)^{-1} X^T y = diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1}) X^T y \rightarrow w_j = (diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1}) X^T y)_j =$$
$$(diag((x_1^T x_1)^{-1}, ...(x_m^T x_m)^{-1}))_j (X^T y)_j = (x_j^T x_j)^{-1} (X^T y)_j = \frac{(X^T y)_j}{x_j^T x_j} = \frac{(X^T)_j y}{x_j^T x_j} = \frac{x_j^T y}{x_j^T x_j}$$

Combining the recent equality with the previous part, results to concluding that ¨optimal parameters from training the regressor on all features is 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 $\|\hat{x_i} - \hat{x_j}\| = \|z_i - z_j\|$.

Solution:

This is just change of basis. Follows from

$$(\hat{x}_i - \hat{x}_j)^T(\hat{x}_i - \hat{x}_j) = (z_i - z_j)^T V_{1:k}^T V_{1:k}(z_i - z_j) = (z_i - z_j)^T(z_i - z_j)$$

since the columns of $V_{1:k}$ are orthogonal.

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.

Solution:

Suppose V is the full $p \times p$ matrix containing all p eigenvectors. Denote by $\tilde{V} = V_{k+1:p}$ the matrix consisting of all eigenvectors except the first k. Since V is orthogonal, $VV^T = I = V_{1:k}V_{1:k}^T + \tilde{V}\tilde{V}^T$.

$$\begin{aligned}
\sum_{i=1}^n \|x_i - \hat{x}_i\|_2^2 &= \sum_{i=1}^n \left\|x_i - V_{1:k}V_{1:k}^\top x_i\right\|_2^2 = \sum_{i=1}^n \left\|\left(I - V_{1:k}V_{1:k}^\top\right) x_i\right\|_2^2 = \sum_{i=1}^n \left\|\tilde{V}\tilde{V}^\top x_i\right\|_2^2 \\
&= \sum_{i=1}^n x_i^\top \tilde{V}\tilde{V}^\top \tilde{V}\tilde{V}^\top x_i = \sum_{i=1}^n x_i^\top \tilde{V}\tilde{V}^\top x_i \\
&= \sum_{i=1}^n \mathrm{Tr}\left[x_i^\top \tilde{V}\tilde{V}^\top x_i\right] = \sum_{i=1}^n \mathrm{Tr}\left[\tilde{V}^\top x_i x_i^\top \tilde{V}\right] = (n-1)\,\mathrm{Tr}\left[\tilde{V}^\top S\tilde{V}\right] \\
&= (n-1)\sum_{j=k+1}^p \lambda_j,
\end{aligned}$$

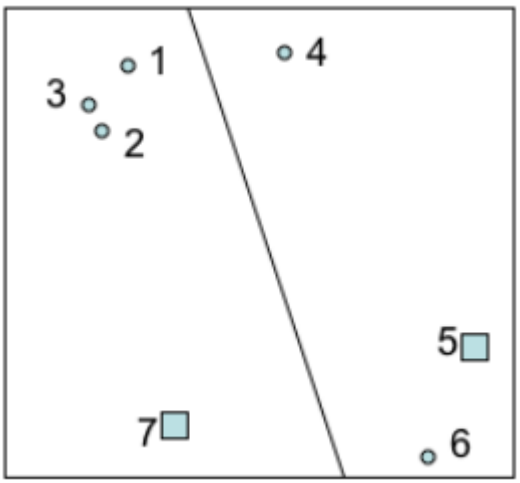
where in the last step we used that $S\tilde{V} = \tilde{V}\Lambda_{k+1:p}$, with $\Lambda_{k+1:p} = \mathrm{diag}\,(\lambda_{k+1}, \lambda_{k+2}, ..., \lambda_p)$.

3 K-means

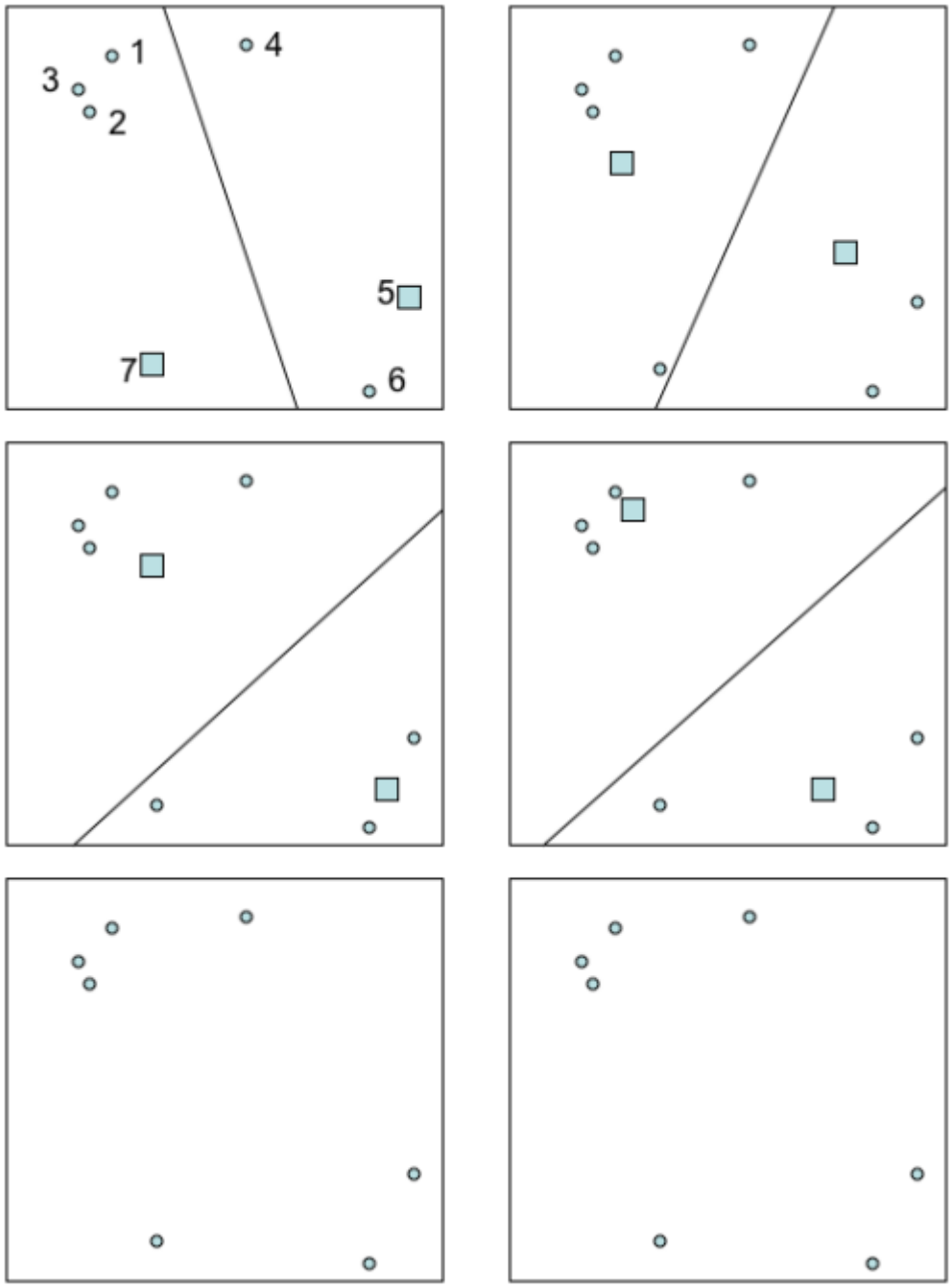
Perform K-means on the dataset given below. Circles are data points and there are two initial cluster centers, at data points shown with squares.

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.



Solution:



3.2

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

Solution:

advantages of hierarchical clustering over K-means:

- Easy to interpret hierarchy for particular applications
- Don't need to know how many clusters you're after

advantages of K-means over hierarchical clustering :

- Can be much faster than hierarchical clustering, depending on data
- Can incorporate new data and reform clusters easily

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

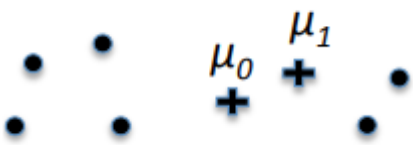
Consider the set of training data below, and two clustering algorithms: K-Means, and GMM using EM (Expectation Maximization). Will these algorithms produce the same cluster centers?

Solution:

Either algorithm will find the clusters just fine. But the difference lies in that k-means uses hard assignment of each point to a single cluster, whereas GMM uses soft assignment, where every point has non-zero (though possibly small) probability of being in each cluster.

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?

Solution:

μ_0 moves to the left, and μ_1 moves to the right.

4.2.2

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

Solution:

Increase. Each iteration of the EM algorithm increases to likelihood of the data, unless you happen to be exactly at a local optimum.