Tyler Olivieri HW 6 #1

Consider a classification problem with N different classes.

let prior probability of class $n \in N$ be π_n

denote $f_n(x) = Pr(X=x|Y=n)$

observations in class n are drawn from $N(\mu_n, \Sigma_n)$ where $\Sigma_n = \Sigma \;\; \forall n$.

a) Use bayes theorem to find $Pr(Y=n|X=x)$

$$Pr(X,Y) = Pr(X,Y)$$

$$Pr(X|y=n)\,p(y=n) = Pr(y|X=x)\,Pr(X=x)$$

$$Pr(y|X=x) = \frac{p(X|y=n)\,p(y=n)}{pr(X=x)}$$

$$\Rightarrow p(y=n|X=x) = \frac{p(X=x|y=n)\,p(y=n)}{p(X=x)}$$

$$= \frac{f_n(x)\,\pi_n}{p(x=x)}$$

$$= \frac{f_n(x)\,\pi_n}{\sum_{i=1}^{k} p(x=x, y=i)}$$

$$= \frac{f_n(x)\,\pi_n}{\sum_{i=1}^{k} p(x=x|y=i)\,p(y=i)}$$

$$P(y=n|X=x) = \frac{f_n(x)\,\pi_n}{\sum^{k} p(x=x|y=i)\,\pi_i}$$

$$\Pr(y=n \mid X=x) = \frac{f_n(x)\,\pi_n}{\sum_{i=1}^{K} f_i(x)\,\pi_i}$$

b) Derive the linear descriminant function, $\delta_n(x)$ and write the classification rule for the predicted class, $\hat{y}$ for an LDA in terms of $\delta_n(x)$.

$$f_n(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\tfrac{1}{2}(x-\mu_n)\Sigma^{-1}(x-\mu_n)\right)$$

$$\Pr(y=n \mid X=x) = \frac{\frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left(-\tfrac{1}{2}(x-\mu_n)\Sigma^{-1}(x-\mu_n)\right)\pi_n}{p(X=x)}$$

$$\log \Pr(y=n \mid X=x) = \log \pi_n + \log\left[\frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left(-\tfrac{1}{2}(x-\mu_n)^T\Sigma^{-1}(x-\mu_n)\right)\right] - \log p(X=x)$$

$$= \log \pi_n + \log\left(\frac{1}{\sqrt{(2\pi)^p|\Sigma|}}\right) + \log\left(\exp\left(-\tfrac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_n - \tfrac{1}{2}\mu_n^T\Sigma^{-1}\mu_n\right)\right) - \log p(X=x)$$

$$\delta_n(x) = \log \Pr(y=n \mid X=x) = \log \pi_n - \tfrac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_n - \tfrac{1}{2}\mu_n^T\Sigma^{-1}\mu_n + \underbrace{\log\left(\frac{1}{\sqrt{(2\pi)^p|\Sigma|}}\right) - \log p(X=x)}_{\text{will drop out}}$$

the classification rule should be

$\hat{y} = \underset{y \in \{1, \dots, n\}}{\text{argmax}} \; \delta_n(x)$

taking the max wrt y removes constants

in $\delta_n(x)$.

$$\hat{y} = \underset{y \in \{1, \dots, n\}}{\text{argmax}} \left(\log \pi_n + x^T \Sigma^{-1} \mu_n - \tfrac{1}{2} \mu_n^T \Sigma^{-1} \mu_n \right)$$

where we estimate π_n, μ_n with $\hat{\pi}_n, \hat{\mu}_n$

c) Derive the decision boundary for the LDA in the case of two classes, a and b.

The decision boundary for two classes a, b is the set of points where

$$\delta_a(x) = \delta_b(x)$$

$$\log \pi_a + x^T \Sigma^{-1} \mu_a - \tfrac{1}{2} \mu_a^T \Sigma^{-1} \mu_a = \log \pi_b + x^T \Sigma^{-1} \mu_b - \tfrac{1}{2} \mu_b^T \Sigma^{-1} \mu_b$$

This is an equation of a line in x.

(substitute estimates for $\pi_a, \pi_b, \mu_a, \mu_b$)

$\hat{\pi}_a, \hat{\pi}_b, \hat{\mu}_a, \hat{\mu}_b$

d) Consider two classes, $a=1$ and $b=2$. You are given $\hat{\pi}_a = .6$, $\hat{\pi}_b = .4$, $\hat{\mu}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\hat{\mu}_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$ and

$$\hat{\Sigma} = \begin{bmatrix} 1.5 & .1 \\ .1 & 1 \end{bmatrix}$$

Find the decision boundary and classification rule of the corresponding LDA. How would the observation $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ be classified?

The decision boundary are the points x s.t

$$\log(.6) + x^T \begin{bmatrix} 1.5 & .1 \\ .1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \tfrac{1}{2} \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1.5 & .1 \\ .1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$= \log(.4) + x^T \begin{bmatrix} 1.5 & .1 \\ .1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} -2 \\ 1 \end{bmatrix} - \tfrac{1}{2} \begin{bmatrix} -2 & 1 \end{bmatrix} \begin{bmatrix} 1.5 & .1 \\ .1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

$$-.51 + x^T \begin{bmatrix} .71 \\ -.07 \end{bmatrix} - .36 = -.92 + x^T \begin{bmatrix} -2.14 \\ 1.21 \end{bmatrix} - 2.75$$

$$-.87 + x^T \begin{bmatrix} .71 \\ -.07 \end{bmatrix} = -3.67 + x^T \begin{bmatrix} -2.14 \\ 1.21 \end{bmatrix}$$

clasify x as class a if

$$-.87 + x^T \begin{bmatrix} .71 \\ -.07 \end{bmatrix} > -3.67 + x^T \begin{bmatrix} -2.14 \\ 1.21 \end{bmatrix}$$

and b otherwise.

$$-.87 + \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} .71 \\ -.07 \end{bmatrix} = -.23$$

$$-3.67 + \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} -2.14 \\ 1.21 \end{bmatrix} = -4.6$$

since $-.23 > -4.6 \Rightarrow$ classify X as a.

e) This relates to logistic regression under the binary case in the following:

with two classes the decision boundary for gaussian bayes is when $\log p(y=1|x) = \log p(y=0|x)$

so choose class 1 when $\log p(y=1|x) - \log p(y=0|x) > 0$

$$\Rightarrow \log \frac{p(y=1|x)}{p(y=0|x)} > 0$$

the quantity $\log \frac{p(y=1|x)}{p(y=0|x)}$ is the log odds

which is the basis of logistic regression

c) The naive Bayes classifier is similar to LDA, except it assumes each predictor is conditionally independent of every other predictor given class n. Derive the classification rule for $\hat{y}$ under this classifier. How does this relate to logistic regression in the binary case (i.e for two classes?)

$$p(X|y) = \prod_{i=1}^{p} p(x_i|y)$$ under naive bayes conditionally independent assumption.

$$p(y=n|X=x) = \frac{f_n(x)\,\pi_n}{p(X=x)} = \frac{\prod_{i=1}^{p} p(x_i|y=n)\,\pi_n}{p(X=x)}$$

$$\log p(y=n|X=x) = \log\left(\prod_{i=1}^{p} p(x_i|y=n)\right) + \log\pi_n - \log p(X=x)$$

$$= \sum_{i=1}^{p} \log p(x_i|y=n) + \log\pi_n + \log p(X=x)$$

$$= \sum_{i=1}^{p} \log \frac{1}{\sqrt{(2\pi)^p}\,\sigma_n^2} \exp\left(-\frac{1}{2\sigma_n^2}(x_i-\mu_n)^2\right) + \log\pi_n + \underbrace{\log p(X=x)}_{\text{constant wrt } y}$$

$$\hat{y} = \underset{y\in\{1,2,\dots,n\}}{\operatorname{argmax}} \log p(y=n|X=x)$$ is classification rule (substitute above expression)

Substitute above expression except for $\log p(X=x)$ as it does not effect maximization wrt y. Substitue estimates for necessary parameters.

Tyler Olivieri HW6 #2 k-means clustering

a) Show that setting the objective ~~fto function~~ the sum of the squared Euclidean distances of points from the center of their cluster

$$\text{obj} = \sum_{k=1}^{K} \sum_{x \in C_k} \sum_{i=1}^{p} (c_{ki} - x_i)^2$$

results in an update rule where the optimal centroid is the mean of the points in the cluster.

min obj wrt c_k

$$\frac{d\,\text{obj}}{d c_k} = \sum_{x \in C_k} \sum_{i=1}^{p} 2(c_{ki} - x_i) = 0$$

$$2 \sum_{x \in C_k} \sum_{i=1}^{p} (c_{ki} - x_i) = 0$$

$$2 \sum_{x \in C_k} \sum_{i=1}^{p} (c_{ki} - x_i) = 0$$

$$\sum_{x \in C_k} \sum_{i=1}^{p} c_{ki} - \sum_{x \in C_k} \sum_{i=1}^{p} x_i = 0$$

$$\sum_{x \in C_k} \sum_{i=1}^{p} c_{ki} = \sum_{x \in C_k} \sum_{i=1}^{p} x_i$$

let n_k be # points assigned to cluster k. $x \in C_k$

$$n_k \sum_{i=1}^{p} c_{ki} = \sum_{x \in C_k} \sum_{i=1}^{p} x_i$$

$$\sum_{i=1}^{p} C_{ki} = \frac{\sum_{x \in C_k} \left(\sum_{i=1}^{p} X_i \right)}{n_k} = \frac{\sum_{i=1}^{p} \sum_{x \in C_k} X_i}{n_k}$$

n_k ← total #'s of points assigned to cluster k.

So set C_{ki} to be the mean of all data points assigned to cluster k.

$$C_{ki} = \frac{\sum_{x \in C_k} X_i}{n_k}$$

for each feature, set cluster center to be mean of data points assigned to the cluster center.

b) Show that setting the objective function to the sum of the manhattan distances of points from the center of their clusters,

$$obj = \sum_{k=1}^{K} \sum_{x \in C_k} \sum_{i=1}^{P} |C_{Ki} - X_i|$$

results in an update rule where the optimal centroid is the median of the cluster.

$$\frac{d\,obj}{d C_K} = \sum_{x \in C_K} \sum_{i=1}^{P} \frac{C_{Ki} - X_i}{|C_{Ki} - X_i|} = 0$$

when $C_{Ki} - X_i > 0$ then $\frac{C_{Ki} - X_i}{|C_{Ki} - X_i|} = \frac{C_{Ki} - X_i}{C_{Ki} - X_i} = 1$

when $C_{Ki} - X_i < 0$ then $\frac{C_{Ki} - X_i}{|C_{Ki} - X_i|} = \frac{C_{Ki} - X_i}{X_i - C_{Ki}} = \frac{C_{Ki} - X_i}{-1(C_{Ki} - X_i)}$

$$= \frac{1}{-1} = -1$$

$\Rightarrow \frac{C_{Ki} - X_i}{|C_{Ki} - X_i|} = \text{sign}(C_{Ki} - X_i)$ where $\text{sign}(x) = 1$ when $x > 0$, $= -1$ when $x < 0$

$$\sum_{i=1}^{P} \left(\sum_{x \in C_K} \text{sign}(C_{Ki} - X_i) \right) = 0$$

This can only equal zero when the number of the positive elements equals the number of negative elements. So for C_{Ki}

$C_{Ki} = \text{median}(X_i \in C_{Ki})$ then

Tyler Olivieri HW6 P3 PCA

Consider the dataset

X	y
0	1
1	1
2	1
2	3
3	2
3	3
4	5

a) Normalize the data and derive the two principal components in sorted order.

$$\mu_x = \frac{0+1+2+2+3+3+4}{7} = 2.14$$

$$\mu_y = \frac{1+1+1+3+2+3+5}{7} = 2.29$$

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)^2 = \frac{1}{6}\sum_{i=1}^{7}(x_i-2.14)^2 = 1.55$$

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{1.55} = 1.24$$

$$\sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\mu_y)^2 = \frac{1}{6}\sum_{i=1}^{7}(y_i-2.29)^2 = 1.92$$

$$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{1.92} = 1.38$$

Normalized data $x_i := \frac{x_i-\mu_x}{\sigma_x}$ $\quad y_i := \frac{y_i-\mu_y}{\sigma_y}$

0 mean, std-dev = 1

X-norm ~ (0,1)
Y-norm " "

X_{norm}	y_{norm}
-1.72	-.93
-.92	-.93
-.11	-.93
-.11	.52
.69	-.21
.69	.52
1.49	1.96

let

$$A = \begin{bmatrix} -1.72 & -.93 \\ -.92 & -.93 \\ -.11 & -.93 \\ -.11 & .52 \\ .69 & .21 \\ .69 & .52 \\ 1.49 & 1.96 \end{bmatrix}$$

be the data matrix.

Find principal components

$T = AW$ where the columns of W are the eigenvectors of A^TA.

(by SVD) $= U\Sigma W^T W$ since W is chosen to be orthonormal $W^TW = I$

$= U\Sigma$

SVD of A: (Done in python)

$$A = \underbrace{\begin{bmatrix} -.52 & -.54 \\ -.36 & .06 \\ -.2 & .56 \\ .08 & -.42 \\ .18 & .33 \\ .24 & .11 \\ .68 & -.32 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 3.6 & 0 \\ 0 & 1.04 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} .71 & .71 \\ .71 & -.71 \end{bmatrix}}_{W^T}$$

Now we have $T = U\Sigma = \begin{bmatrix} -.52 & -.54 \\ -.36 & .06 \\ -.2 & .56 \\ -.08 & -.42 \\ .18 & .33 \\ .24 & .11 \\ .68 & .32 \end{bmatrix} \begin{bmatrix} 3.6 & 0 \\ 0 & 1.04 \end{bmatrix}$

$$T = U\Sigma = \begin{bmatrix} -1.88 & -.56 \\ -1.31 & .06 \\ -.74 & .58 \\ .29 & -.44 \\ .64 & .34 \\ .86 & .12 \\ 2.44 & -.33 \end{bmatrix}$$

The new transformed dataset using the first principal component is

$$\check{X} = \begin{bmatrix} -1.88 \\ -1.31 \\ -.74 \\ .29 \\ .64 \\ .86 \\ 2.44 \end{bmatrix}$$

Which is the first column of T.

b) repeat the previous analysis but do not normalize the data.

Is pca scale-invariant

SVD of A_unnormalized: (removing ~~n-r~~ columns of U, and 0's rows and columns of Σ)

when you multiply $U\Sigma$, they are irrelevant.

because python libs give SVD this way except numpy, which I did not use

$$A\_unnormalized = \underbrace{\begin{bmatrix} .08 & .44 \\ .15 & -.04 \\ .22 & -.51 \\ .37 & .37 \\ .37 & -.55 \\ .45 & -.11 \\ .67 & .30 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 9.52 & 0 \\ 0 & 1.54 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} .68 & .73 \\ -.73 & .68 \end{bmatrix}}_{V^T}$$

$$T\_unnormalized = U\Sigma = \begin{bmatrix} .73 & .68 \\ 1.41 & -.06 \\ 2.09 & -.79 \\ 3.60 & .56 \\ 3.50 & -.85 \\ 4.24 & -.17 \\ 6.39 & .46 \end{bmatrix}$$

$$\hat{X}\_unnormalized = \begin{bmatrix} .73 \\ 1.41 \\ 2.09 \\ 3.60 \\ 3.50 \\ 4.24 \\ 6.39 \end{bmatrix}$$

$\Rightarrow$ PCA is <u>not</u> scale-invariant $\hat{X} \neq \hat{X}\_unnormalized$

$$T = U\Sigma = \begin{bmatrix} -1.88 & -.56 \\ -1.31 & .06 \\ -.74 & .58 \\ .29 & -.44 \\ .64 & .34 \\ .86 & .12 \\ 2.44 & -.33 \end{bmatrix}$$

The new transformed dataset using the first principal component is

$$\tilde{X} = \begin{bmatrix} -1.88 \\ -1.31 \\ -.74 \\ .29 \\ .64 \\ .86 \\ 2.44 \end{bmatrix}$$

Which is the first column of T.