TA session 3

Peter, Pochien, Sheray

Outline

- Overview on HW2 hand-writing part
- Overview on Midterm

HW2 question 1

- In Principal Components Analysis, given a x and its a low-dimensional projection y with the following equation: $\mathrm{y} = \mathrm{w}^{\mathrm{T}} x$, please answer the following questions:
- (a) Please explain connection between w and covariance matrix of y. [5pt]
- (b) Assume the covariance matrix of x is $\begin{matrix} 16 & 0 & 2 \\ 0 & 25 & 0 \\ 2 & 0 & 4 \end{matrix}$ and we set $\mathrm{k} = 1$ in PCA, please determine the matrix w. [10pt]

HW2 question 1: (a) answer

(a)

- The projection of $\boldsymbol{x}$ on the direction of $\boldsymbol{w}$ is: $z = \boldsymbol{w}^T\boldsymbol{x}$
- Find $\boldsymbol{w}$ such that $\boxed{\mathrm{Var}(z)}$ is maximized

information

$$\begin{aligned}
\mathrm{Var}(z) = \mathrm{Var}(\boldsymbol{w}^T\boldsymbol{x}) &= E[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})^2] \\
&= E[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})] \\
&= E[\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{w}] \\
&= \boldsymbol{w}^T E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T]\boldsymbol{w} = \boldsymbol{w}^T \Sigma \boldsymbol{w}
\end{aligned}$$

where $\mathrm{Var}(\boldsymbol{x}) = E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T] = \Sigma$

HW2 question 1: (b) answer

(b)

- First projection: Maximize Var(z_1) subject to $|| \boldsymbol{w_1} ||=1$

$$\max_{\mathbf{w}_1} \mathbf{w}_1^T \Sigma \mathbf{w}_1 - \alpha\left(\mathbf{w}_1^T \mathbf{w}_1 - 1\right)$$ ← A Lagrange problem

$\boxed{\Sigma \boldsymbol{w}_1 = \alpha \boldsymbol{w}_1}$ that is, $\boldsymbol{w}_1$ is an eigenvector of Σ

Choose the one with the largest eigenvalue for Var(z) to be max

For derive PCA with k=1, you need to compute the biggest eigenvalue and the corresponding eigenvector for covariance matrix of y.

The equation of finding eigenvalues could be written as:

$$\det\left(\begin{pmatrix} 16 & 0 & 2 \\ 0 & 25 & 0 \\ 2 & 0 & 4 \end{pmatrix} - \lambda \mathbf{I}\right) = 0$$, we could get three eigenvalues

$$\lambda = 25, 16.3245, 3.6754$$

the biggest one is 25, and its corresponding eigenvector is [0,1,0]=W

(Remind: This means that the PCA only keep the second feature of X)

HW2 question 2: Answers

- Define a multivariate Bernoulli mixture where inputs are binary and derive the EM equation.[10pt]

When the components are multivariate Bernouilli, we have binary vectors that are d-dimensional. Assuming that the dimensions are independent, we have (see section 5.7)

$$p_i(\mathbf{x}^t|\Phi) = \prod_{j=1}^{d} p_{ij}^{x_j^t}(1-p_{ij})^{1-x_j^t}$$

where $\Phi^l = \{p_{i1}^l, p_{i2}^l, \ldots, p_{id}^l\}_{i=1}^k$. The E-step does not change (equation 7.9). In the M-step, for the component parameters $p_{ij}, i = 1, \ldots, k, j = 1, \ldots, d$, we maximize

$$\begin{aligned} Q' &= \sum_t \sum_i h_i^t \log p_i(\mathbf{x}^t|\phi^l) \\ &= \sum_t \sum_i h_i^t \sum_j x_j^t \log p_{ij}^l + (1-x_j^t)\log(1-p_{ij}^l) \end{aligned}$$

Taking the derivative with respect to p_{ij} and setting it equal to 0, we get

$$p_{ij}^{l+1} = \frac{\sum_t h_i^t x_j^t}{\sum_t h_j^t}$$

Note that this is the same as in equation 5.31, except that estimated "soft" labels h_i^t replace the supervised labels r_i^t.

HW2 question 3: Answers

- Generalize the Gini index and the misclassification error for K > 2 classes. Generalize misclassification error to risk, taking a loss function into account.[5pt]

  - Gini index with $K > 2$ classes: $\phi(p_1, p_2, \ldots, p_K) = \sum_{i=1}^{K} \sum_{j<i} p_i p_j$
  - Misclassification error: $\phi(p_1, p_2, \ldots, p_K) = 1 - \max_{i=1}^{K} p_i$
  - Risk: $\phi_\Lambda(p_1, p_2, \ldots, p_K) = \min_{i=1}^{K} \sum_{k=1}^{K} \lambda_{ik} p_k$ where Λ is the $K \times K$ loss matrix (equation 3.7).

HW2 question 4: Answers

- Show that the derivative of the softmax, $\mathrm{y_i} = \frac{\exp(\mathrm{a_i})}{\sum_j \exp(a_j)}$, is $\frac{\partial y_i}{\partial a_j} = y_i(\delta_{ij} - y_j)$ where δ_{ij} is 1 if $\mathrm{i} = \mathrm{j}$ and 0 otherwise. [10pt]

Given that

$$y_i = \frac{\exp a_i}{\sum_j \exp a_j}$$

for $i = j$, we have

$$\begin{aligned}
\frac{\partial y_i}{\partial a_i} &= \frac{\exp a_i \left(\sum_j \exp a_j\right) - \exp a_i \exp a_j}{\left(\sum_j \exp a_j\right)^2} \\
&= \frac{\exp a_i}{\sum_j \exp a_j}\left(\frac{\sum_j \exp a_j - \exp a_i}{\sum_j \exp a_j}\right) \\
&= y_i(1 - y_i)
\end{aligned}$$

and for $i \neq j$, we have

$$\begin{aligned}
\frac{\partial y_i}{\partial a_j} &= \frac{-\exp a_i \exp a_j}{\left(\sum_j \exp a_j\right)^2} \\
&= -\left(\frac{\exp a_i}{\sum_j \exp a_j}\right)\left(\frac{\sum_j \exp a_j}{\sum_j \exp a_j}\right) \\
&= y_i(0 - y_j)
\end{aligned}$$

which we can combine in one equation as

$$\frac{\partial y_i}{\partial a_j} = y_i(\delta_{ij} - y_j)$$

Overview on Midterm

- Grade distribution:

Grade range	Members	Percentage
110	0	0%
100 ~ 109	5	7%
90 ~ 99	5	7%
80 ~ 89	11	17%
70 ~ 79	20	31%
60 ~ 69	9	14%
50 ~ 59	5	7%
40 ~ 49	6	9%
30 ~ 39	2	3%
20 ~ 29	0	0%
10 ~ 19	0	0%
0 ~ 9	0	0%

Mean	73.14
Std	17.09
Median	74

Midterm Q4 (a), Ans

[35pt] Given a data set:

$$X = \{\vec{x}_0, \vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\} \text{ ,where } \vec{x}_i \in R^{1\times 5}.$$

Please answer the following questions:

(a) [10pt, 3+3+4] Assume all the pairwise correlation coefficient of features of $\vec{\mathrm{x}}_\mathrm{i}$ is 0, the mean and variance in each feature are given as follow:

$$\vec{\mu} = [2, 5, 14, -4, 5],$$
$$\sigma^2 = [0.25, 16, 64, 0.04, 81]$$

We intend to perform a dimensional reduction using a Principal Components Analysis (PCA) with dimension parameter k = 2. A PCA dimension reduction on the input data matrix X to obtain the reduced data matrix Y can be rewritten using the following matrix form:

$$Y = W^T X \qquad (eq1)$$

Find W and explain briefly what W means and does.

1. PCA k=2 → reduce dimension from 5 to 2
2. How?

   Correlation coefficient = 0

   → Covariance = 0

   → Keep the highest variance features

   → Keep the second highest variance features

   → The fifth feature (81) and the third feature (64)
3. W^T =[0,0,0,0,1,

   0,0,1,0,0]

Midterm Q4 (b) (c), Ans

(b) [5pt] Following (a), we further adapt $eq1$ to build a binary classification using linear discrimination method by applying a sigmoid function σ:

$$g(X) = \sigma(W^T X)$$

where $g(X)$ is the discriminant function for binary classification. Assume another W_θ for X using PCA with $k = 1$, and there is another dataset $Z = \{\vec{z}_0, \vec{z}_1, \vec{z}_2, \vec{z}_3\}$ shown as follows:

$$\vec{z}_0 = [1, 0.5, 8, 15, 2]$$
$$\vec{z}_1 = [1, -0.2, 9, 1, -2]$$
$$\vec{z}_2 = [1, 0.7, 0.7, 14, 2]$$
$$\vec{z}_3 = [1, -0.6, 8, 2, -2]$$

What are the predicted classes for dataset Z after W_θ projection for this classifier? (e = 2.71)

[Hint: $P(X) \in C_0$ $if\ P(X) < 0.5$, $P(X) \in C_1$ $if\ P(X) > 0.5$]

(c) [5pt] Following (b), is using W_θ projection reasonable for reducing dimensions on dataset Z in constructing this classifier? Please explain in detail.

1. PCA k=1 for X set

→ reduce dimension from 5 to 1

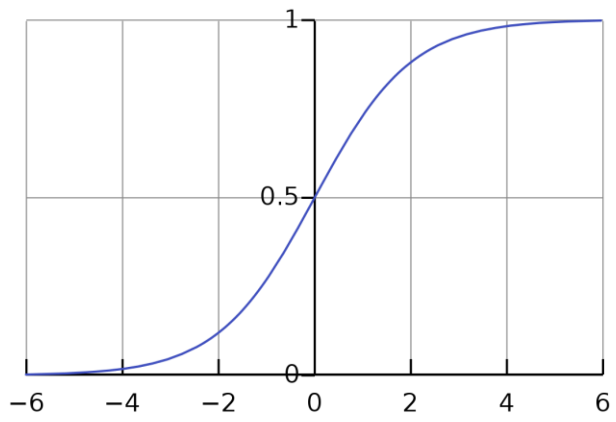
2. How?

→ Keep the highest variance features

→ The fifth feature (81)

→ Extract the fifth feature in Z set

3. Pass sigmoid function for classification



4. $\vec{z}_0 \in C_1, \vec{z}_1 \in C_0, \vec{z}_2 \in C_1, \vec{z}_3 \in C_0$

5. Not reasonable: feature of Z datasets has its own distribution, PCA should be done on Z's features, but not X's features.

(Partial point for explain a lot, even if your answer is wrong)

Midterm Q4 (d), Ans

[5pt] Given n data points X and label L:

$$L = \{r_0, r_1, r_2, \dots, r_n\},$$
$$\text{where } r_i = 0 \text{ if } \vec{x}_i \in C_1 \text{ and } r_i = 1 \text{ if } \vec{x}_i \in C_2$$

Here, we intend to perform Linear Discriminant Analysis (LDA) for dimensional reduction on X using a matrix W_α. $\mathrm{M_j} \in R^5$ *and* $\mathrm{m_j} \in R^1$ denote the mean vectors of samples and means of samples after W_α projection for j^{th} class, respectively, where $j \in \{1, 2\}$.

Please write the formula for the scatter of samples s_1^2 and s_2^2 from C_1 and C_2 after projection.

Linear Discriminant Analysis

- Find a low-dimensional space s.t. when ***x*** is projected classes are as well separated as possible (supervised method)
  - $z = W^T x$: projection of x onto w, a dimensional from d -> 1
- **m_1: original sample mean,** m_1: after projection
- **m_2: original sample mean,** m_2: after projection
- For a sample, $X = \{x^t, r^t\}$, $r^t = 1$ is x^t=class 1

$$m_1 = \frac{\sum_t \mathbf{w}^T \mathbf{x}^t r^t}{\sum_t r^t} \qquad s_1^2 = \sum_t \left(\mathbf{w}^T \mathbf{x}^t - m_1\right)^2 r^t$$

After projection for m1, s1 (scatter)

- ind ***w*** that maximizes

10

$$s_1^2 = \sum_t (w^T x^t - m_1)^2 r^t$$

$$s_2^2 = \sum_t (w^T x^t - m_2)^2 (1 - r^t)$$

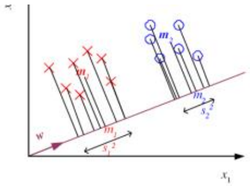
Score:

If you don't use $(1 - r^t)$, you should reveal the corresponding $\vec{x}_i \in C_1$ or $\vec{x}_i \in C_2$.

Midterm Q4 (e), Ans

(e) [10pt] Follow (d), find W_α using Fisher's linear discriminant.

LDA



- The goal is to let mean as far apart as possible and let the scatter for each class to be as clustered as possible
- $|m1\text{-}m2|^2$ large, s_1^2 + $s_2$2 as small

Fisher's discriminant, find w, such that

$$\boxed{J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}}$$

Is maximized

Detail: 6.41-6.46

- Between-class scatter:

$$\begin{aligned}(m_1 - m_2)^2 &= \left(\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2\right)^2 \\ &= \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} \\ &= \mathbf{w}^T\mathbf{S}_B\mathbf{w} \text{ where } \mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\end{aligned}$$

- Within-class scatter:

$$\begin{aligned}s_1^2 &= \sum_t \left(\mathbf{w}^T\mathbf{x}^t - m_1\right)^2 r^t \\ &= \sum_t \mathbf{w}^T(\mathbf{x}^t - \mathbf{m}_1)(\mathbf{x}^t - \mathbf{m}_1)^T\mathbf{w}r^t = \mathbf{w}^T\mathbf{S}_1\mathbf{w}\end{aligned}$$

$$\text{where } \mathbf{S}_1 = \sum_t (\mathbf{x}^t - \mathbf{m}_1)(\mathbf{x}^t - \mathbf{m}_1)^T r^t$$

$$s_1^2 + s_1^2 = \mathbf{w}^T\mathbf{S}_W\mathbf{w} \text{ where } \mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

12

Fisher's Linear Discriminant

- Find *w* that max

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}} = \frac{\left|\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)\right|^2}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}$$

- LDA soln: $\boxed{\mathbf{w} = c \cdot \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)}$

Only direction matters, set c=1

Score:

One red box: 6 pt

If you write both without explaining: 8 pt

Midterm Q5 (a),(b), Ans

5. [20pt] If assume Gaussian component (each mixture is a Gaussian distribution), derive the **M-step** equations for:
   (a) [10pt] **S**, in the case of shared **arbitrary covariance matrix**
   (hint: $p_i(x^t|\theta) \sim N(\boldsymbol{m}_i, \boldsymbol{S})$)
   (b) [10pt] s^2, in the case of shared **diagonal covariance matrix**
   (hint: $p_i(x^t|\theta) \sim N(\boldsymbol{m}_i, s^2\boldsymbol{I})$)

(a)
M-step:

$$\Phi^{l+1} = \arg\max_{\Phi} \mathcal{Q}(\Phi|\Phi^l)$$

$$\begin{aligned} \mathcal{Q}(\Phi|\Phi^l) &= \sum_t \sum_i h_i^t [\log \pi_i + \log p_i(\boldsymbol{x}^t|\Phi)] \\ &= \sum_t \sum_i h_i^t \log \pi_i + \boxed{\sum_t \sum_i h_i^t \log p_i(\boldsymbol{x}^t|\Phi)} \end{aligned}$$

Detail: 7.10 – 7.16
Score:
If I feel you tried to write something but you fail → 5 pt

In the case of a shared arbitrary covariance matrix, in the E-step, we have

$$h_i^t = \frac{\pi_i \exp[-(1/2)(\boldsymbol{x}^t - \boldsymbol{m}_i)^T \mathbf{S}_i^{-1}(\boldsymbol{x}^t - \boldsymbol{m}_i)]}{\sum_j \pi_j \exp[-(1/2)(\boldsymbol{x}^t - \boldsymbol{m}_j)^T \mathbf{S}_j^{-1}(\boldsymbol{x}^t - \boldsymbol{m}_j)]}$$

and in the M-step for the component parameters, we have

$$\min_{\boldsymbol{m}_i, \mathbf{S}} \sum_t \sum_i h_i^t (\boldsymbol{x}^t - \boldsymbol{m}_i)^T \mathbf{S}^{-1}(\boldsymbol{x}^t - \boldsymbol{m}_i)$$

The update equation for $\boldsymbol{m}_i$ does not change but for the common covariance matrix, we have

$$\begin{aligned} \mathbf{S}^l &= \frac{\sum_t \sum_i h_i^t (\boldsymbol{x}^t - \boldsymbol{m}_i^{l+1})(\boldsymbol{x}^t - \boldsymbol{m}_i^{l+1})^T}{\sum_t \sum_i h_i^t} \\ &= \frac{\sum_t \sum_i h_i^t (\boldsymbol{x}^t - \boldsymbol{m}_i^{l+1})(\boldsymbol{x}^t - \boldsymbol{m}_i^{l+1})^T}{N} \end{aligned}$$

Another way we can see this is by considering that

$$\mathbf{S} = \sum_i P(\mathcal{G}_i)\mathbf{S}_i = \sum_i \left(\frac{\sum_t h_i^t}{N}\right)\mathbf{S}_i$$

In the case of a shared diagonal matrix, for the E-step we have

$$h_i^t = \frac{\pi_i \exp\left[-(1/2s^2)\|\boldsymbol{x}^t - \boldsymbol{m}_i\|^2\right]}{\sum_j \pi_j \exp\left[-(1/2s^2)\|\boldsymbol{x}^t - \boldsymbol{m}_j\|^2\right]}$$

and in the M-step, we have

$$\min_{\boldsymbol{m}_i, s} \sum_t \sum_i h_i^t \frac{\|\boldsymbol{x}^t - \boldsymbol{m}_i\|^2}{s^2}$$

The update equation for the shared variance is

$$s^2 = \frac{\sum_t \sum_i \sum_{k=1}^d h_i^t (x_k^t - m_{ik})^2}{Nd}$$

Midterm Q6 ,Ans

6. (Bonus)[10pt] Consider a density model given by a mixture distribution

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|k)$$

and suppose that we partition the vector x into two parts so that $x = (x_a, x_b)$. Show that the conditional density $p(x_b|x_a)$ is itself a mixture distribution and find expressions for the mixing coefficients and for the component densities.

Problem 9.10 Solution

According to the property of PDF, we know that:

$$p(\mathbf{x}_b|\mathbf{x}_a) = \frac{p(\mathbf{x}_a, \mathbf{x}_b)}{p(\mathbf{x}_a)} = \frac{p(\mathbf{x})}{p(\mathbf{x}_a)} = \sum_{k=1}^{K} \frac{\pi_k}{p(\mathbf{x}_a)} \cdot p(\mathbf{x}|k)$$

Note that here $p(\mathbf{x}_a)$ can be viewed as a normalization constant used to guarantee that the integration of $p(\mathbf{x}_b|\mathbf{x}_a)$ equal to 1. Moreover, similarly, we can also obtain:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \sum_{k=1}^{K} \frac{\pi_k}{p(\mathbf{x}_b)} \cdot p(\mathbf{x}|k)$$