CS182: Introduction to Machine Learning
Reference Solutions of Final Exam

June 12, 2022

I Regression and Probability Estimation [12 points]

We consider the following linear regression model in which y is the sum of a deterministic linear function of x, plus random noise ϵ, i.e.,

$$y = wx + \epsilon, \tag{1}$$

where x is the real-valued input, y is the real-valued output, and w is a single real-valued parameter to be learned. Here ϵ is a real-valued random variable that represents noise which follows a Gaussian distribution with mean 0 and standard deviation σ, that is, $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
Note: the probability density function $f(X)$ of a Gaussion distributed variable $X \sim \mathcal{N}(\mu, \sigma^2)$ takes the form

$$f(X = x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}). \tag{2}$$

1. **[4 points]** Write down the probability distribution of y conditioned on x and w., i.e. $\Pr(y \mid w, x)$.

Solution

$$\Pr(y \mid w, x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y - wx)^2}{2\sigma^2}).$$

2. **[4 points]** Given n *i.i.d.* training examples $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. Let $\mathcal{Y} = (y_1, ..., y_n)$ and $\mathcal{X} = (x_1, ..., x_n)$, please write down an expression for the conditional data likelihood: $\Pr(\mathcal{Y} \mid \mathcal{X}, w)$

Solution

$$\begin{aligned}
\Pr(\mathcal{Y} \mid \mathcal{X}, w) &= \prod_{i=1}^{n} \Pr(y_i \mid x_i, w) \\
&= (\frac{1}{2\pi\sigma^2})^{n/2} \prod_{i=1}^{n} \exp(-\frac{(y_i - wx_i)^2}{2\sigma^2}) \\
&= (\frac{1}{2\pi\sigma^2})^{n/2} \exp(-\frac{\sum_{i=1}^{n}(y_i - wx_i)^2}{2\sigma^2}).
\end{aligned}$$

3. **[4 points]** Suppose a Laplace prior over w with $\mu = 0$ and b (i.e., $w \sim \mathit{Laplace}(0, b)$). Now you need to use MAP(maximum a posterior probability) to estimate w from the training data. Please show that finding the MAP estimate w^* is equivalent to solving the following optimization problem

$$w^* = \underset{w}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - wx_i)^2 + c|w|. \tag{3}$$

Express the regularization parameter c in terms of σ and b.
Hint: the probability density function $f(X)$ of a Laplace distributed variable $X \sim \mathit{Laplace}(\mu, b)$ takes the form

$$f(X = x) = \frac{1}{2b} \exp(-\frac{|x - \mu|}{b}). \tag{4}$$

Solution

$$
\begin{aligned}
\Pr(w \mid \mathcal{Y},\ \mathcal{X}) &\propto \Pr(\mathcal{Y} \mid \mathcal{X}, w)\Pr(w) \\
&\propto \exp(-\frac{\sum_{i=1}^{n}(y_i - wx_i)^2}{2\sigma^2})\exp(-\frac{|w|}{b}) \\
w^* &= \operatorname*{argmin}_w\ -\ln\Pr(w \mid \mathcal{Y},\ \mathcal{X}) \\
&= \operatorname*{argmin}_w \frac{\sum_{i=1}^{n}(y_i - wx_i)^2}{2\sigma^2} + \frac{|w|}{b} \\
&= \operatorname*{argmin}_w \frac{1}{2}\sum_{i=1}^{n}(y_i - wx_i)^2 + \frac{\sigma^2}{b}|w|.
\end{aligned}
$$

We can find that $c = \frac{\sigma^2}{b}$.

II Linear Classification [12 points]

Let X be a d-dimensional binary vector, drawn from one of two classes: P or Q. Assume each element X_i in X is an independent Bernoulli random variable with parameter p_i when X drawn from class P (similarly, with parameter q_i for class Q). That is

$$\begin{aligned} X_i|(Y = P) &\sim Bernoulli(p_i), \qquad 1 \le i \le d, \\ X_i|(Y = Q) &\sim Bernoulli(q_i), \qquad 1 \le i \le d. \end{aligned}$$

Note: for this problem, the values of p_i and q_i, along with priors $\Pr(Y = P) = \pi_p$ and $\Pr(Y = P) = \pi_q$, are known.

1. [**3 points**] Given a vector $x \in \{0, 1\}^d$, compute the probabilities $\Pr(X = x|Y = P)$ and $\Pr(X = x|Y = Q)$ in terms of class parameters p_i and q_i. Your answer must be a single expression for each probability.

Solution

$$\Pr(X = x|Y = P) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i},$$
$$\Pr(X = x|Y = Q) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}.$$

2. [**4 points**] Please write down the equation which holds if and only if x is at the decision boundary of the Bayes' optimal classifier.

Solution

$$\Pr(Y = P|X = x) = \frac{\Pr(X = x|Y = P)\Pr(Y = P)}{\Pr(X = x)},$$
$$\Pr(Y = Q|X = x) = \frac{\Pr(X = x|Y = Q)\Pr(Y = Q)}{\Pr(X = x)},$$

Therefore, the equation of decision boundary is

$$\pi_p \Pr(X = x|Y = P) = \pi_q \Pr(X = x|Y = Q).$$

3. [**5 points**] The decision boundary derived above is actually linear in x, which can be expressed as:

$$\{x \in \{0, 1\}^d | w^T x + b = 0\},$$

for some vector w and scalar b. Please find expressions for w and b in terms of priors (π_p and π_q) and class parameters (p_i and q_i).

Solution

$$\Pr(Y = P)\Pr(X = x|Y = P) = \Pr(Y = Q)\Pr(X = x|Y = Q)$$
$$\pi_p \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i} = \pi_q \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}$$
$$ln(\pi_p) + \sum_{i=1}^{d}[x_i ln(p_i) + (1 - x_i)ln(1 - p_i)] = ln(\pi_q) + \sum_{i=1}^{d}[x_i ln(q_i) + (1 - x_i)ln(1 - q_i)]$$

where we can get:

$$\sum_{i=1}^{d}[(ln\frac{p_i}{q_i} - ln\frac{1-p_i}{1-q_i})x_i] + ln\frac{\pi_p}{\pi_q} + \sum_{i=1}^{d} ln\frac{1-p_i}{1-q_i} = 0.$$

Therefore,

$$w_i = ln\frac{p_i}{q_i} - ln\frac{1-p_i}{1-q_i}$$

$$b = ln\frac{\pi_p}{\pi_q} + \sum_{i=1}^{d} ln\frac{1-p_i}{1-q_i}.$$

III Graphical Model [12 points]

We have a Bayesian network shown below, in which $X_1, X_2, ..., X_8$ are eight boolean random variables. Please answer the following questions.
Note: correct answers without proof will get 0 point.

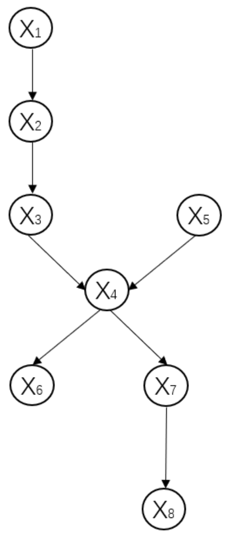

Figure 1: The Bayesian network with eight variables.

1. **[3 points]** Now we have known probabilities for some random variables. For X_1, we have $\Pr(x_1) = 0.7$. For X_2, we have $\Pr(x_2|x_1) = 0.6$ and $\Pr(x_2|\neg x_1) = 0.3$. For X_3, we have $\Pr(x_3|x_2) = 0.4$ and $\Pr(x_3|\neg x_2) = 0.8$. Apply the method of inference to calculate marginal probability $\Pr(\neg x_3)$.
Note: please round your results to 3 decimal places.

Solution

$$\begin{aligned}\Pr(\neg x_3) &= \sum_{x_1, x_2} \Pr(x_1, x_2, \neg x_3) \\ &= \Pr(x_1)\Pr(x_2|x_1)\Pr(\neg x_3|x_2) + \Pr(x_1)\Pr(\neg x_2|x_1)\Pr(\neg x_3|\neg x_2) \\ &\quad + \Pr(\neg x_1)\Pr(x_2|\neg x_1)\Pr(\neg x_3|x_2) + \Pr(\neg x_1)\Pr(\neg x_2|\neg x_1)\Pr(\neg x_3|\neg x_2) \\ &= 0.7 \times 0.6 \times 0.6 + 0.7 \times 0.4 \times 0.2 + 0.3 \times 0.3 \times 0.6 + 0.3 \times 0.7 \times 0.2 \\ &= 0.404.\end{aligned}$$

2. **[3 points]** Using the same probabilities for X_1, X_2 X_3 in III.1, and apply the method of inference to calculate conditional probability $\Pr(\neg x_2|\neg x_3)$.
Note: please round your results to 3 decimal places.

Solution

$$\begin{aligned}\Pr(\neg x_2, \neg x_3) &= \sum_{x_1} \Pr(x_1, \neg x_2, \neg x_3) \\ &= \Pr(x_1)\Pr(\neg x_2|x_1)\Pr(\neg x_3|\neg x_2) + \Pr(\neg x_1)\Pr(\neg x_2|\neg x_1)\Pr(\neg x_3|\neg x_2) \\ &= 0.7 \times 0.4 \times 0.2 + 0.3 \times 0.7 \times 0.2 \\ &= 0.098.\end{aligned}$$

So $\Pr(\neg x_2|\neg x_3) = \frac{\Pr(\neg x_2, \neg x_3)}{\Pr(\neg x_3)} = 0.243.$

3. **[3 points]** Prove that $X_1 \perp\!\!\!\perp X_3 | X_2$ without using D-separation.

> **Solution**
> $\Pr(X_1, X_3|X_2) = \frac{\Pr(X_1,X_2,X_3)}{\Pr(X_2)} = \frac{\Pr(X_1,X_2)\Pr(X_3|X_2)}{\Pr(X_2)} = \Pr(X_1|X_2)\Pr(X_3|X_2)$.

4. **[3 points]** Discuss whether the statement, $X_1 \perp\!\!\!\perp X_5 | X_6$, is true or not, and explain the reason based on D-separation.

> **Solution**
> The statement is false.
> Given X_6, the path through X_1 to X_5 is open.

IV Expectation-Maximization [10 points]

Given a Bayesian network with four discrete variables $\{A, B, C, D\}$, where $\{A, C, D\}$ are boolean variables and $B \in \{0, 1, 2\}$. Suppose that $\{A, C, D\}$ are observed variables and $\{B\}$ is a latent variable. Now we implement EM algorithm for this model. Suppose there are K observations in total. ($\{a_k, c_k, d_k\}_{k=1}^{K}$).

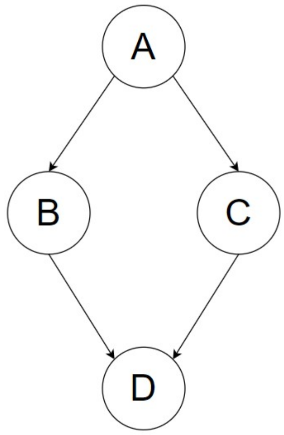

Figure 2: The Bayesian network with four discrete variables $\{A, B, C, D\}$.

1. [**4 points**] Derive the E-step.

Solution

In E-step, calculate $P(B|A, C, D, \theta)$.

$$P(b_k = 0|a_k, c_k, d_k, \theta) = \frac{P(b_k = 0, a_k, c_k, d_k|\theta)}{\sum_{i=0}^{2} P(b_k = i, a_k, c_k, d_k|\theta)},$$
$$P(b_k = 1|a_k, c_k, d_k, \theta) = \frac{P(b_k = 1, a_k, c_k, d_k|\theta)}{\sum_{i=0}^{2} P(b_k = i, a_k, c_k, d_k|\theta)},$$
$$P(b_k = 2|a_k, c_k, d_k, \theta) = \frac{P(b_k = 2, a_k, c_k, d_k|\theta)}{\sum_{i=0}^{2} P(b_k = i, a_k, c_k, d_k|\theta)}.$$

2. [**6 points**] Derive the M-step, and update parameters for the Bayesian network

Solution

In M-step, choose θ' which maximize $E_{P(B|A,C,D,\theta)} \log P(A, B, C, D|\theta')$, where

$$\begin{aligned} &E_{P(B|A,C,D,\theta)} \log P(A, B, C, D|\theta') \\ &= \sum_{k=1}^{K} \sum_{i=0}^{2} P(b_k = i|a_k, c_k, d_k, \theta)[\log P(a_k) + \log P(b_k|a_k) + \log P(c_k|a_k) + \log P(d_k|b_k, c_k)]. \end{aligned}$$

Parameters are updated based on:

$$
\begin{aligned}
\theta_a &= \frac{\sum_{k=1}^{K} \delta(a_k = 1)}{K}, \\
\theta_{b|a} &= \frac{\sum_{k=1}^{K} P(b_k = b)\delta(a_k = a)}{\sum_{k=1}^{K} \delta(a_k = a)}, \\
\theta_{c|a} &= \frac{\sum_{k=1}^{K} \delta(a_k = a, c_k = 1)}{\sum_{k=1}^{K} P(a_k = a)}, \\
\theta_{d|b,c} &= \frac{\sum_{k=1}^{K} \delta(d_k = 1, c_k = c)P(b_k = b)}{\sum_{k=1}^{K} \delta(c_k = c)P(b_k = b)}.
\end{aligned}
$$

V Support Vector Machines [12 points]

Support vector machines (SVM) are supervised learning models, that directly optimize for the maximum margin separator. Fig. 3 shows an example of maximum margin separator over a dataset $S = \{(x_i, y_i)\}_{i=1}^n$, in which $x_i \in \mathbb{R}^2$ and $y_i \in \{-1, 1\}$ denote the i-th sample and the i-th label ($\forall i$), respectively. For simplicity, here we assume that the dataset S has been standardized, and thus the bias can be omitted in the linear model. In Fig. 3, "+" and "-" denote the samples with labels "1" and "-1", respectively, and $\mathbf{w}$ is the normal vector of the maximum margin separator $\mathbf{w}^\top x = 0$. You need to derive the optimization problem of SVM in the linearly separable case.
Note: correctly giving the results without detailed derivation will get 0 point.

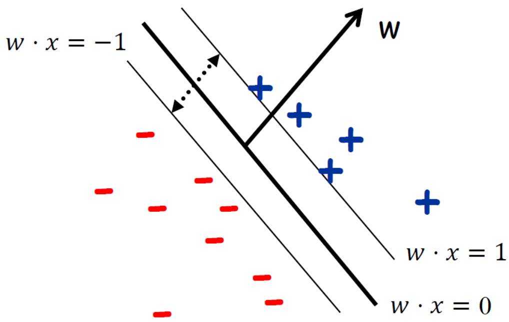

Figure 3: Maximum margin separator in the linearly separable case.

1. **[5 points]** Derive the constraint optimization problem of SVM in the separable case shown in Fig. 3.

Solution
Let r be the margin between $\mathbf{w}^\top x = 0$ and $\mathbf{w}^\top x = 1$. Assume there are two points $x_0 \in \mathbb{R}^2$ and $x_1 \in \mathbb{R}^2$ on $\mathbf{w}^\top x = 0$ and $\mathbf{w}^\top x = 1$, respectively, and we make $x_1 - x_0$ paralleled with $\mathbf{w}$. Hence, we have the following equations:

$$\begin{cases} w^\top x_1 = 1, \\ w^\top x_0 = 0, \\ x_1 - x_0 = r \times \frac{\mathbf{w}}{||\mathbf{w}||_2}, \end{cases}$$

where $||\cdot||_2$ denotes the ℓ_2-norm. By multiplying $\mathbf{w}^\top$ on both sides of the third equation, and plugging the first two equations into it, we have

$$\begin{aligned} \mathbf{w}^\top (x_1 - x_0) &= r \times \frac{\mathbf{w}^\top \mathbf{w}}{||\mathbf{w}||_2} \\ 1 &= r \times ||\mathbf{w}||_2, \\ \Rightarrow \quad r &= \frac{1}{||\mathbf{w}||_2}. \end{aligned}$$

In the separable case, a maximum margin separator should satisfy the following three conditions:

- maximize the margin $r = \frac{1}{||\mathbf{w}||_2}$ over a dataset;
- put positive samples ($y_i = 1$) on one side of the separator, i.e., $\mathbf{w}^\top x_i \geq 1$;
- put negative samples ($y_i = -1$) on another side of the separator, i.e., $\mathbf{w}^\top x_i \leq -1$.

Therefore, the constraint optimization problem of SVM is

$$\begin{aligned} &\min_{\mathbf{w}} \; ||\mathbf{w}||_2^2, \\ &\text{s.t. } y_i \mathbf{w}^\top x_i \geq 1, \; \forall i \in \{1, 2, ..., n\}. \end{aligned}$$

2. **[5 points]** Derive the dual problem of the above primal problem based on K.K.T. conditions.

Solution

We first formulate the Lagrangian function of the primal problem:

$$L(\mathbf{w}, \xi, \alpha, \lambda) = ||\mathbf{w}||_2^2 - \sum_{i=1}^{n} \alpha_i (y_i \mathbf{w}^\top x_i - 1),$$

where $\alpha_i \geq 0$ $(\forall i)$ is the dual variable. Because strong duality holds in the primal problem, the optimal optimization variables $\{\mathbf{w}^*, \alpha^*\}$ should satisfy K.K.T. conditions:

- primal: $y_i {\mathbf{w}^*}^\top x_i \geq 1$, $\forall i$,
- dual: $\alpha_i^* \geq 0$, $\forall i$,
- complementary: $\alpha_i^* (y_i {\mathbf{w}^*}^\top x_i - 1) = 0,\ \forall i$,
- stationary: $\nabla_{\mathbf{w}^*} L = 0$.

According to the stationary condition, we have

$$\nabla_{\mathbf{w}} L = 2\mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i x_i = 0, \quad \Rightarrow \quad \mathbf{w} = \frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i x_i,$$

Substituting them into the Lagrangian function yields the dual function $g(\alpha, \lambda)$,

$$\begin{aligned}
g(\alpha) &= \inf_{\mathbf{w}, \xi} L(\mathbf{w}, \alpha) \\
&= \frac{1}{4} \langle \sum_{i=1}^{n} \alpha_i y_i x_i, \sum_{j=1}^{n} \alpha_j y_j x_j \rangle - \sum_{i=1}^{n} \alpha_i (y_i \langle \frac{1}{2} \sum_{j=1}^{n} \alpha_j y_j x_j, x_i \rangle - 1) \\
&= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i.
\end{aligned}$$

Thus, the dual problem is

$$\begin{aligned}
&\max_{\alpha} \ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i, \\
&\text{s.t. } \alpha_i \geq 0,\ \forall i.
\end{aligned}$$

3. **[2 points]** Kernel functions implicitly define some mapping function $\phi(\cdot)$ that transforms an input instance $x \in \mathbb{R}^d$ to a high or even infinite dimensional feature space Q, by giving the form of dot product in $Q : k(x_i, x_j) = \phi(x_i) \dot{\phi}(x_j)$. Please kernelize the dual problem, in order to learn a non-linear SVM classifier.

Solution

Based on the dual problem, the kernelized SVM problem becomes

$$\begin{aligned}
&\max_{\alpha} \ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) + \sum_{i=1}^{n} \alpha_i, \\
&\text{s.t. } \alpha_i \geq 0,\ \forall i.
\end{aligned}$$

VI Clustering [10 points]

Given six data points in 2D space (shown in Table 1) and two initial cluster centers $c_1 = (0, 1), c_2 = (0, -1)$, please answer the following questions.

i	x	y
1	-2	1
2	0	2
3	2	1
4	2	-1
5	0	-2
6	-2	-1

Table 1: Six input data points

1. [**5 points**] Please use k-means algorithm to cluster the given points into two groups.

Solution
The first iteration is shown at Table 2.
According to the table, it is obvious that x_1, x_2 and x_3 should be clustered into one group and x_4, x_5 and x_6 should be clustered into the other group. The center points for two groups are $c_1^{new} = (0, \frac{4}{3})$ and $c_2^{new} = (0, -\frac{4}{3})$.

i	x	y	distance to c_1	distance to c_2
1	-2	1	2	$2\sqrt{2}$
2	0	2	1	3
3	2	1	2	$2\sqrt{2}$
4	2	-1	$2\sqrt{2}$	2
5	0	-2	3	1
6	-2	-1	$2\sqrt{2}$	2

Table 2: Results of the first iteration.

2. [**5 points**] Please give the center points for the two groups after the algorithm converges.

Solution
The second iteration is shown at Table 3.
We can find that the clustering result keeps the same as the first iteration so the algorithm converges. Above all, x_1, x_2 and x_3 should be clustered into one group with the center point $(0, \frac{4}{3})$ and x_4, x_5 and x_6 should be clustered into the other group with center point $(0, -\frac{4}{3})$.

i	x	y	distance to c_1^{new}	distance to c_2^{new}
1	-2	1	$\sqrt{37}/3$	$\sqrt{85}/3$
2	0	2	$2/3$	$10/3$
3	2	1	$\sqrt{37}/3$	$\sqrt{85}/3$
4	2	-1	$\sqrt{85}/3$	$\sqrt{37}/3$
5	0	-2	$10/3$	$2/3$
6	-2	-1	$\sqrt{85}/3$	$\sqrt{37}/3$

Table 3: Results of the second iteration.

VII Dimensionality Reduction [12 points]

Given three data points in 2D space: $x_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, x_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, please answer the following questions:

Note: correct answers without detailed derivation will get 0 point.

1. [**4 points**] What are the first and second principal components?

Solution

$\mathbf{X} = [x_1, x_2, x_3] = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}, \mathbf{X}^T = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{bmatrix}, \mathbf{XX}^T = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

$\mathbf{XX}^T - \lambda\mathbf{I} = \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix}, |\mathbf{XX}^T - \lambda\mathbf{I}| = (2-\lambda)^2 - 1 = 0 \Rightarrow (\lambda-3)(\lambda-1) = 0 \Rightarrow \lambda_1 = 3, \lambda_2 = 1$

When $\lambda = 3, \mathbf{XX}^T - \lambda\mathbf{I} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \Rightarrow \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix} \Rightarrow v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

When $\lambda = 1, \mathbf{XX}^T - \lambda\mathbf{I} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \Rightarrow v_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

So the first and second principal component directions are $v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, v_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ respectively.

2. [**4 points**] If we project the original data points on the new coordinate system represented by the principal components, what are their coordinates?

Solution

Let z_1, z_2, z_3 denote the points in the new coordinate system represented by the principal component directions.

$z_1 = \begin{bmatrix} x_1^T v_1 \\ x_1^T v_2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, z_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}, z_3 = \begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix}.$

3. [**4 points**] What is the variance of the data in each direction? Verify that it is equal to the total variance of the origin data.

Solution

Variance of the first direction: $Var_1 = \frac{1}{3}[(-\frac{1}{\sqrt{2}})^2 + (-\frac{1}{\sqrt{2}})^2 + (\sqrt{2})^2] = 1.$

Variance of the second direction: $Var_2 = \frac{1}{3}[(\frac{1}{\sqrt{2}})^2 + (-\frac{1}{\sqrt{2}})^2] = \frac{1}{3}.$

Total variance of the origin data: $Var_{origin} = \frac{1}{3}[(-1)^2 + 1^2 + (-1)^2 + 1^2] = \frac{4}{3}.$

It is obvious that $Var_{origin} = Var_1 + Var_2$.

VIII Neural Networks [12 points]

As shown in Fig.4, we have a feed-forward neural network with two hidden-layer nodes and one output node, and x_1 and x_2 are two inputs. For simplicity, the bias b is omitted here. For the following questions, assume the learning rate η in gradient descent is fixed by $\eta = 0.1$. Both hidden and output units use the same activation function $g(\cdot)$.

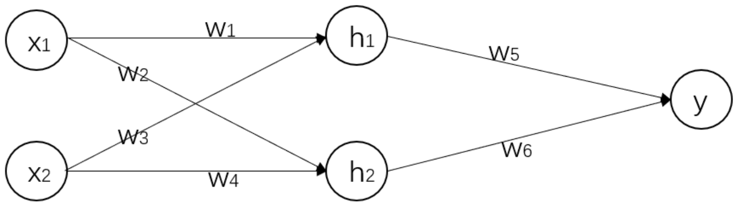

Figure 4: The Neural network with one hidden layer.

1. **[4 points]** Express the output y_{output} in terms of inputs x_1, x_2, weights $w_1, w_2, w_3, w_4, w_5, w_6$ and the activation function g.

Solution

$$y_{\text{output}} = g(w_5 h_1 + w_6 h_2) = g(w_5 g(w_1 x_1 + w_3 x_2) + w_6 g(w_2 x_1 + w_4 x_2)).$$

2. **[8 points]** Assume we have one input $\{x_1 = 1, x_2 = 1\}$ and the real target of it is $y_{\text{target}} = 1$. The initial value of $w_1^{(0)}, w_2^{(0)}, w_3^{(0)}, w_4^{(0)}, w_5^{(0)}, w_6^{(0)}$ is 1,2,-1,$\frac{1}{2}$,-2,1. And the loss on the given example is defined as $L = \frac{1}{2}(y_{\text{target}} - y_{\text{output}})^2$. Suppose that the sigmoid activation function $g(z) = \frac{1}{1+e^{-z}}$ is used.
Note: please round your results to 3 decimal places.

(1) **[3 points]** Without any optimization, calculate the output h_1, h_2 and y_{output} on the given example.

Solution

$$\begin{aligned}
h_1 &= g(w_1 x_1 + w_3 x_2) = g(0) = \frac{1}{2} \\
h_2 &= g(w_2 x_1 + w_4 x_2) = g(2.5) = 0.924 \\
y_{\text{output}} &= g(w_5 g(w_1 x_1 + w_3 x_2) + w_6 g(w_2 x_1 + w_4 x_2)) \\
&= g(-2g(0) + g(2.5)) \\
&= 0.481.
\end{aligned}$$

(2) **[5 points]** Compute the updated weights $w_1^{(1)}, w_2^{(1)}, w_3^{(1)}, w_4^{(1)}, w_5^{(1)}, w_6^{(1)}$ by performing ONE step of gradient descent. Show all steps in your calculation.

Solution

$$
\begin{aligned}
&\Delta w_5 = (0.481 - 1) \times 0.481 \times (1 - 0.481) \times 0.5 = -0.0648,\\
&\Delta w_6 = (0.481 - 1) \times 0.481 \times (1 - 0.481) \times 0.924 = -0.1198,\\
&\Delta w_1 = (0.481 - 1) \times 0.481 \times (1 - 0.481) \times (-2) \times 0.5 \times (1 - 0.5) \times 1 = 0.0648,\\
&\Delta w_2 = (0.481 - 1) \times 0.481 \times (1 - 0.481) \times (1) \times 0.924 \times (1 - 0.924) \times 1 = -0.0091,\\
&\Delta w_3 = (0.481 - 1) \times 0.481 \times (1 - 0.481) \times (-2) \times 0.5 \times (1 - 0.5) \times 1 = 0.0648,\\
&\Delta w_4 = (0.481 - 1) \times 0.481 \times (1 - 0.481) \times (1) \times 0.924 \times (1 - 0.924) \times 1 = -0.0091,\\
&w_1^{(1)} = w_1^{(0)} - 0.1 \times \Delta w_1 = 0.994,\\
&w_2^{(1)} = w_2^{(0)} - 0.1 \times \Delta w_2 = 2.001,\\
&w_3^{(1)} = w_3^{(0)} - 0.1 \times \Delta w_3 = -1.006,\\
&w_4^{(1)} = w_4^{(0)} - 0.1 \times \Delta w_4 = 0.501,\\
&w_5^{(1)} = w_5^{(0)} - 0.1 \times \Delta w_5 = -1.994,\\
&w_6^{(1)} = w_6^{(0)} - 0.1 \times \Delta w_6 = 1.012.
\end{aligned}
$$

IX Convolutional Neural Networks [8 points]

Convolutional neural networks are designed to process 2D features instead of the 1D ones in multi-layer perceptron (MLP).

1. [**4 points**] Please calculate the feature map based on 2D convolution, if you are given the following 5×5 image matrix in Table 4 and 2×2 kernel matrix in Table 5. (stride = 1, no padding)

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

Table 4: 5×5 image matrix.

1	0
0	1

Table 5: 2×2 kernel matrix.

Solution
The feature map is shown in Table 6.

8	10	12	14
18	20	22	24
28	30	32	34
38	40	42	44

Table 6: 4×4-feature maps.

2. [**4 points**] Based on the above result, calculate the feature maps after max-pooling and average-pooling, respectively. (both pooling with 2×2 filters and stride = 2)

Solution
Please refer to Tables 7 and 8.

20	24
40	44

Table 7: After max-pooling.

14	18
34	38

Table 8: After average-pooling.