SI151: Optimization and Machine Learning
Final Exam Solutions
(June 29, 2020)

1. (12 points) *Regression function and least squares.*
Given the input variables $X \in \mathbb{R}^d$ and response variable $Y \in \mathbb{R}$, the Expected Prediction Error (EPE) is defined by

$$\text{EPE}(f) = \mathbb{E}[L(Y, f(X))], \tag{1}$$

where $\mathbb{E}(\cdot)$ denotes the expectation over the joint distribution $\Pr(X, Y)$, and $L(Y, f(X))$ is the squared error loss function

$$L(Y, f(X)) = (Y - f(X))^2, \tag{2}$$

measuring the difference between the observed Y and estimated function $f(X)$.

   (a) [5pts] Based on the assumption of linearity, we can approximate $f(X)$ by a linear model $X^\top\beta$. Please derive the linear estimator β by minimizing $\text{EPE}(f)$ w.r.t. β.

   (b) [5pts] Given a dataset $\{(x_i, y_i)\}_{i=1}^n$, where x_i and y_i denote the i-th sample and the i-th label ($\forall i$), respectively, please derive the least squares (LS) estimator by minimizing the residual sum of squares (RSS):

$$\text{RSS}(\beta) = \sum_{i=1}^{n}(y_i - x_i^\top\beta)^2 = \|\mathbf{y} - \mathbf{X}\beta\|_2^2, \tag{3}$$

   where $\|\cdot\|_2$ denotes the ℓ_2 norm, $\mathbf{y} = [y_1, y_2, ..., y_n]^\top \in \mathbb{R}^n$, and $\mathbf{X} = [x_1, x_2, ..., x_n]^\top \in \mathbb{R}^{n\times d}$ with full column rank.

   (c) [2pts] Explain how the LS estimator in (b) approximates the linear estimator in (a).

★ **Solution**:

(a) Plugging $f(X) = X^\top\beta$ into $\text{EPE}(f)$ yields

$$\begin{aligned}
\text{EPE}(f) =& \mathbb{E}\left((Y - X^\top\beta)^2\right) \quad \text{(1 point)} \\
=& \mathbb{E}\left((Y - X^\top\beta)^\top(Y - X^\top\beta)\right) \\
=& \mathbb{E}\left(Y^\top Y - Y^\top X^\top\beta - \beta^\top XY + \beta^\top XX^\top\beta\right) \\
=& \mathbb{E}(Y^\top Y) - \mathbb{E}^\top(XY)\beta - \beta^\top\mathbb{E}(XY) + \beta^\top\mathbb{E}(XX^\top)\beta, \quad \text{(2 points)}
\end{aligned}$$

which is a convex quadratic function w.r.t. β, and must achieve its minimum when $\frac{\partial\text{EPE}(f)}{\partial\beta} = 0$.

$$\begin{aligned}
&\frac{\partial\text{EPE}(f)}{\partial\beta} = -2\mathbb{E}(XY) + 2\mathbb{E}(XX^\top)\beta = 0, \\
&\Rightarrow \quad \beta = \left(\mathbb{E}(XX^\top)\right)^{-1}\mathbb{E}(XY). \quad \text{(2 points)}
\end{aligned}$$

(b) Given an arbitrary vectors a, we have the following equality:

$$\|a\|_2^2 = \langle a, a\rangle = a^\top a.$$

Thus, $\text{RSS}(\beta)$ is reformulated by

$$\begin{aligned}
\text{RSS}(\beta) &= \langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta\rangle \\
&= \langle \mathbf{y}, \mathbf{y}\rangle - 2\langle \mathbf{y}, \mathbf{X}\beta\rangle + \langle \mathbf{X}\beta, \mathbf{X}\beta\rangle \\
&= \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\beta + \beta^\top\mathbf{X}^\top\mathbf{X}\beta. \quad \text{(2 points)}
\end{aligned}$$

Since $\mathrm{RSS}(\beta)$ is a convex function, setting its derivative w.r.t. β as 0 leads to the optimal solution.

$$\frac{\partial \mathrm{RSS}(\beta)}{\partial \beta} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\beta = 0,$$
$$\Rightarrow \quad \hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}, \quad \text{(2 points)}$$

where the second equality holds because the data matrix $\mathbf{X}$ has full column rank. (1 point)

(c) The LS estimator,

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$
$$= \left(\frac{\mathbf{X}^\top \mathbf{X}}{n-1}\right)^{-1} \frac{\mathbf{X}^\top \mathbf{y}}{n-1},$$

approximates the linear regression estimator $\beta = \left(\mathbb{E}(XX^\top)\right)^{-1} \mathbb{E}(XY)$ by replacing the theoretical expectation with unbiased estimation over the observed data. (2 points)

2. (12 points) *Linear regression and parameter estimation.*
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.,

$$y = wx + \epsilon, \tag{4}$$

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, and 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 Gaussian distributed variable $X \sim \mathcal{N}(\mu, \sigma^2)$ takes the form

$$f(X = x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{5}$$

where μ and σ^2 denote mean and variance, respectively.

(a) [2pts] Write down the probability distribution of y conditioned on x and w.

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

(c) [6pts] The primary target of this question is for you to derive the expression for obtaining a MAP (maximum a posterior probability) estimate of w from the training data. Suppose a Gaussian prior over w with mean 0 and standard deviation τ (i.e., $w \sim \mathcal{N}(0, \tau^2)$). Please show that finding the MAP estimate w^* is equivalent to solving the following optimization problem

$$w^* = \arg\min_w \frac{1}{2}\sum_{i=1}^{n}(y_i - wx_i)^2 + \frac{\lambda}{2}w^2. \tag{6}$$

Also explicitly express the regularization parameter λ in terms of σ and τ.

★ **Solution:**

(a) The output y follows a Gaussian distribution with the mean wx and the standard deviation σ:

$$\Pr(y \mid w, x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y - wx)^2}{2\sigma^2}\right\} \quad \text{(2 points)}$$

(b)

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

(c)

$$\begin{aligned}
\Pr(w|\mathcal{Y}, \mathcal{X}) &\propto \Pr(\mathcal{Y}|\mathcal{X}, w)\Pr(w|\mathcal{X}) \\
&\propto \Pr(\mathcal{Y}|\mathcal{X}, w)\Pr(w) \quad \text{(1 point)} \\
&\propto \exp\left\{-\frac{\sum_{i=1}^{n}(y_i - wx_i)^2}{2\sigma^2}\right\} \exp\left\{-\frac{w^2}{2\tau^2}\right\} \quad \text{(1 point)} \\
w^* &= \text{argmax}_w \ln \Pr(w|\mathcal{Y}, \mathcal{X}) \quad \text{(2 points)} \\
&= \text{argmax}_w -\frac{\sum_{i=1}^{n}(y_i - wx_i)^2}{2\sigma^2} - \frac{w^2}{2\tau^2} \\
&= \text{argmin}_w \frac{\sum_{i=1}^{n}(y_i - wx_i)^2}{2\sigma^2} + \frac{w^2}{2\tau^2} \\
&= \text{argmin}_w \frac{1}{2}\sum_{i=1}^{n}(y_i - wx_i)^2 + \frac{\sigma^2}{2\tau^2}w^2. \quad \text{(1 point)}
\end{aligned}$$

We can see that $\lambda = \frac{\sigma^2}{\tau^2}$. (1 point)

3. (12 points) *Linear classification and decision boundary.*
A binary image is a digital image where each pixel has only possible values: zero (white) or one (black). A binary image, which consists of a grid of pixels, can therefore naturally be represented as a vector with entries in $\{0, 1\}$.

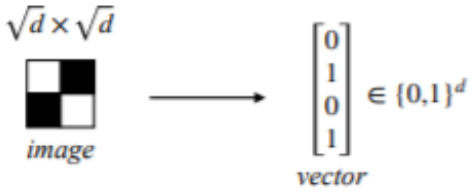


Figure 1: Example of vectorization of a binary image.

In this problem, we consider a classification scheme based on a simple generative model. Let X be a random binary image, represented as a d-dimensional binary vector, drawn from one of two classes: P or Q. Assume every pixel X_i is an independent Bernoulli random variable with parameter p_i and q_i when drawn from classes P and Q respectively.

$$X_i|Y = P \sim \text{Bernoulli}(p_i), \qquad \text{independently for all } 1 \le i \le d, \tag{7}$$
$$X_i|Y = Q \sim \text{Bernoulli}(q_i), \qquad \text{independently for all } 1 \le i \le d. \tag{8}$$

Note: for this problem, we focus on the ideal case, where the true values of p_i and q_i , along with priors π_p and π_q, are known.

(a) [2pts] Given an image $x \in \{0, 1\}^d$, compute the probabilities $\Pr(X = x|Y = P)$ and $\Pr(X = x|Y = Q)$ in terms of the priors, image pixels and/or class parameters. Your answer must be a single expression for each probability.

(b) [4pts] In terms of the probabilities above, write an equation which holds if and only if x is at the decision boundary of the Bayes' optimal classifier. No simplification is necessary for full credit.

(c) [6pts] It turns out that the decision boundary derived above is actually linear in the features of x, so for some vectors w and scalar b, it can be succinctly expressed as:

$$\{x \in \{0, 1\}^d | w^T x + b = 0\}. \tag{9}$$

Find the entries of the vector w and value of b in terms of class priors and parameters.

★ **Solution**:

(a) Because conditional probability of X_i follows Bernoulli distribution, and is independent w.r.t. X_j $(\forall i \neq j)$, we have

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

(b) Decision boundary of Bayes' optimal classifier:

$$\{x \in \{0, 1\}^d | \Pr(Y = P|X = x) = \Pr(Y = Q|X + x)\}. \quad \text{(1 point)}$$

Based on Bayes theorem, we have

$$\Pr(Y = P|X = x) = \frac{\Pr(X = x|Y = P)\Pr(Y = P)}{\Pr(X = x)},$$
$$\Pr(Y = Q|X = x) = \frac{\Pr(X = x|Y = Q)\Pr(Y = Q)}{\Pr(X = x)}. \quad \text{(1 point)}$$

Hence, the equation on the decision boundary becomes

$$\pi_p \Pr(X = x|Y = P) = \pi_q \Pr(X = x|Y = Q). \quad \text{(2 points)}$$

(c)

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

$$\pi_p \prod_{i=1}^{d} p_i^{x_i} (1 - p_i)^{1 - x_i} = \pi_q \prod_{i=1}^{d} q_i^{x_i} (1 - q_i)^{1 - x_i} \quad \text{(1 point)}$$

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

$$\ln \frac{\pi_p}{\pi_q} + \sum_{i=1}^{d} x_i \frac{\ln p_i}{\ln q_i} + \sum_{i=1}^{d} (1 - x_i) \frac{\ln(1 - p_i)}{\ln(1 - q_i)} = 0. \quad \text{(3 points)}$$

So $w_i = \ln \dfrac{p_i}{q_i} - \ln \dfrac{1 - p_i}{1 - q_i}$, $b = \ln \dfrac{\pi_p}{\pi_q} + \sum_{i=1}^{d} \dfrac{\ln(1 - p_i)}{\ln(1 - q_i)}$. (2 points)

4. (12 points) *Representation by Bayesian network.*
The Bayesian network shown in Fig. 2 represents the joint probability distribution of eight boolean random variables, $X_1, X_2, ..., X_8$. Please answer the following questions.
Note: correct answers without proof will get 0 point.

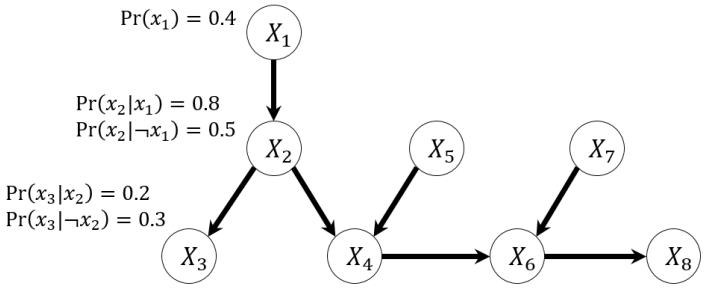


Figure 2: Bayesian network with eight boolean random variables.

(a) [4pts] Apply the method of inference to calculate marginal probability $\Pr(\neg x_3)$.

(b) [4pts] Apply the method of inference to calculate conditional probability $\Pr(x_2 \mid \neg x_3)$.

(c) [2pts] Validate the statement $X_1, X_3 \perp\!\!\!\perp X_7 \mid X_8$.

(d) [2pts] Validate the statement $X_5 \perp\!\!\!\perp X_7 \mid \varnothing$.

★ **Solution:**

(a) When calculating $\Pr(\neg x_3)$ (and $\Pr(x_2 \mid \neg x_3)$ analogically), X_4 is a leaf that is not a query nor evidence, thus it can be eliminated without changing the target probabilities.

$$\begin{aligned}\Pr(\neg x_3) &= \sum_{x_1,x_2} \Pr(x_1, x_2, \neg x_3) = \sum_{x_1,x_2} \Pr(x_1)\Pr(x_2 \mid x_1)\Pr(\neg x_3 \mid x_2) \quad \text{(1 point)} \\ &= \Pr(x_1)Pr(x_2 \mid x_1)\Pr(\neg x_3 \mid x_2) + \Pr(x_1)\Pr(\neg x_2 \mid x_1)\Pr(\neg x_3 \mid \neg x_2) \\ &\quad + \Pr(\neg x_1)\Pr(x_2 \mid \neg x_1)\Pr(\neg x_3 \mid x_2) + \Pr(\neg x_1)\Pr(\neg x_2 \mid \neg x_1)\Pr(\neg x_3 \mid \neg x_2) \quad \text{(1 point)} \\ &= .4 \times .8 \times .8 + .4 \times .2 \times .7 + .6 \times .5 \times .8 + .6 \times .5 \times .7 \quad \text{(1 point)} \\ &= \mathbf{.762} \quad \text{(1 point)}\end{aligned}$$

The same result is reached when editing the following expression:

$$\begin{aligned}\Pr(\neg x_3) &= \sum_{x_1,x_2,x_4} \Pr(x_1, x_2, \neg x_3, x_4) = \sum_{x_1,x_2,x_4} \Pr(x_1)\Pr(x_2 \mid x_1)\Pr(\neg x_3 \mid x_2)\Pr(x_4 \mid x_2) \\ &= \sum_{x_1,x_2} \Pr(x_1)\Pr(x_2 \mid x_1)\Pr(\neg x_3 \mid x_2) \sum_{x_4} \Pr(x_4 \mid x_2) \\ &= \sum_{x_1,x_2} \Pr(x_1)\Pr(x_2 \mid x_1)\Pr(\neg x_3 \mid x_2) \times 1\end{aligned}$$

(b) Analogically, $\Pr(x_2, \neg x_3)$ and $\Pr(x_2 \mid \neg x_3)$ can be calculated by

$$\begin{aligned}\Pr(x_2, \neg x_3) &= \sum_{x_1} \Pr(x_1, x_2, \neg x_3) = \sum_{x_1} \Pr(x_1)\Pr(x_2 \mid x_1)\Pr(\neg x_3 \mid x_2) \quad \text{(1 point)} \\ &= \Pr(x_1)\Pr(x_2 \mid x_1)\Pr(\neg x_3 \mid x_2) + \Pr(\neg x_1)\Pr(x_2 \mid \neg x_1)\Pr(\neg x_3 \mid x_2) \\ &= .4 \times .8 \times .8 + .6 \times .5 \times .8 \\ &= \mathbf{.496} \quad \text{(1 point)}\end{aligned}$$

Then, we have

$$\Pr(x_2 \mid \neg x_3) = \frac{\Pr(x_2, \neg x_3)}{\Pr(\neg x_3)} = \mathbf{.6509} \quad \text{(2 points)} \qquad \approx .651$$

(c) FALSE. (1 point)
Given X_8, the path through X_2, X_4 and X_6 is opened. (1 point)

(d) TRUE. (1 point)
The path through X_4 and X_6 is blocked. (1 point)

5. (10 points) *VC dimension and sample complexity.*
The VC dimension, VC(H), of hypothesis space H defined over instance space $\mathcal{X}$, is the size of the largest number of points (in some configuration) that can be shattered by H. Suppose with probability $(1-\delta)$, a PAC learner outputs a hypothesis within error ϵ of the best possible hypothesis in H. It can be shown that the lower bound on the number of training examples m sufficient for successful learning, stated in terms of VC(H) is

$$m \geq \frac{1}{\epsilon}(4\log_2\frac{2}{\delta} + 8 \times \mathrm{VC}(H)\log_2\frac{13}{\epsilon}). \tag{10}$$

Consider a learning problem in which $\mathcal{X} = \mathbb{R}$ is the set of real numbers, and the hypothesis space is the set of intervals $H = \{(a < x < b) \mid a, b \in \mathbb{R}\}$, which labels points inside the interval as positive, and negative otherwise.
Note: correct answers without proof will get 0 point.

(a) [5pts] What is the VC dimension of H?

(b) [5pts] What is the probability that a hypothesis consistent with m examples will have error at least ϵ?

★ **Solution:**

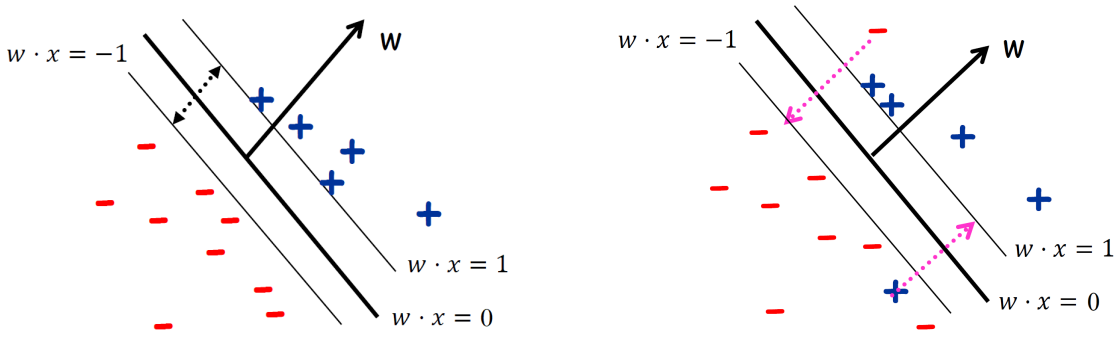
(a) VC(H) = 2. Suppose we have two points x_1 and x_2, and $x_1 < x_2$. They can always be shattered by H, no matter how they are labeled.
- if x_1 positive and x_2 negative, choose $a < x_1 < b < x_2$;
- if x_1 negative and x_2 positive, choose $x_1 < a < x_2 < b$;
- if both x_1 and x_2 positive, choose $a < x_1 < x_2 < b$;
- if both x_1 and x_2 negative, choose $a < b < x_1 < x_2$. (3 points)

However, if we have three points $x_1 < x_2 < x_3$ and if they are labeled as x_1 (positive) x_2 (negative) and x_3 (positive), then they cannot be shattered by H. (2 points)

(b) Substituting VC(H) = 2 into the inequality yields

$$\begin{aligned}
m &\geq \frac{1}{\epsilon}\left(4\log_2\frac{2}{\delta} + 8 \times 2\log_2\frac{13}{\epsilon}\right) \quad \text{(2 points)} \\
\epsilon m &\geq 4\log_2\frac{2}{\delta} + 8 \times 2\log_2\frac{13}{\epsilon} \\
\epsilon m - 16\log_2\frac{13}{\delta} &\geq 4\log_2\frac{2}{\delta} \\
\frac{2^{\frac{\epsilon m}{4}}}{\left(\frac{13}{\epsilon}\right)^4} &\geq \frac{2}{\delta} \\
\delta &\geq \frac{\left(\frac{13}{\epsilon}\right)^4}{2^{\frac{\epsilon m}{4}-1}}. \quad \text{(3 points)}
\end{aligned}$$

6. (16 points) *SVM, duality and kernel methods.*
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, in both separable case and non-separable case. 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 $\mathbf{w}^\top x = 0$. In this problem, you need to derive the linear optimization problem of SVM in both primal and dual forms, and finally extend it for non-linear classification.
Note: correctly giving the results without detailed derivation will get 0 point.



(a) Separable case (b) Non-separable case

Figure 3: Maximum margin separator.

(a) [4pts] Derive the constraint optimization problem of SVM in the separable case shown in Fig. 3(a).

(b) [2pts] Extend the results in (a) to handle the non-separable case shown in Fig. 3(b).

(c) [3pts] Determine the convexity of the problem in (b), and explain whether strong duality holds.

(d) [5pts] Derive the dual problem of the original problem in (b) based on K.K.T. conditions.

(e) [2pts] Extend the linear model in (d) for non-linear classification by the kernel trick.

★ **Solution:**

(a) 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:

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

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

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

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., $\mathbf{w}^\top x_i \leq -1$. (1 point)

Therefore, the constraint optimization problem of SVM is

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

(b) To handle the non-separable case shown in Fig. 3(b), we need introduce the slack variable $\xi_i \geq 0$ $(\forall i)$ to move the possible outlier x_i to the correct side of the separator, and penalize the total amount of ξ_i in the objective function. (1 point)
Thus, the optimization problem becomes

$$\begin{aligned} \min_{\mathbf{w},\xi} \ & \|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i, \\ \text{s.t. } & y_i\mathbf{w}^\top x_i \geq 1-\xi_i, \ \forall i, \\ & \xi_i \geq 0, \forall i, \quad \text{(1 point)} \end{aligned}$$

where $C > 0$ is the regularization parameter.

(c) The problem in (b) is a convex optimization problem due to the following two reasons.

- The objective function $f_0(\mathbf{w},\xi)$ of the optimization problem in (b) is

$$f_0(\mathbf{w},\xi) = \|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i.$$

  Since $\|\cdot\|_2^2$ is squared ℓ_2-norm, the first term of $f_0(\mathbf{w},\xi)$ is convex. The second term of $f_0(\mathbf{w},\xi)$ is a non-negative summation of linear functions, therefore, it is convex. Based on the same reason and $C > 0$, we can conclude that $f_0(\mathbf{w},\xi)$ is convex. (1 point)
- The constrains of the problem in (b) are all linear inequalities. (1 point)

For convex optimization problem with all constraints being linear inequalities, Slater's condition is always satisfied once the solution is feasible. Thus, strong duality holds in the problem of (b). (1 point)

(d) We first formulate the Lagrangian function of the problem of (b):

$$L(\mathbf{w},\xi,\alpha,\lambda) = \|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i\mathbf{w}^\top x_i - 1 + \xi_i) - \sum_{i=1}^{n}\lambda_i\xi_i, \quad \text{(1 point)}$$

where $\alpha_i \geq 0$ and $\lambda_i \geq 0$ $(\forall i)$ are dual variables. Because strong duality holds in the problem of (b), the optimal optimization variables $\{\mathbf{w}^*,\xi^*,\alpha^*,\lambda^*\}$ should satisfy K.K.T. conditions:

- primal: $y_i{\mathbf{w}^*}^\top x_i \geq 1-\xi_i^*$, $\xi_i^* \geq 0$, $\forall i$,
- dual: $\alpha_i^* \geq 0$, $\lambda_i^* \geq 0$, $\forall i$,
- complementary:

$$\begin{aligned} \alpha_i^*(y_i{\mathbf{w}^*}^\top x_i - 1 + \xi_i^*) = 0, \ \forall i, \\ \lambda_i^*\xi_i^* = 0, \ \forall i. \end{aligned}$$

- stationary: $\nabla_{\mathbf{w}^*}L = \nabla_{\xi^*}L = 0$. (1 point)

According to the stationary condition, we have

$$\begin{aligned} \nabla_{\mathbf{w}}L = 2\mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i x_i = 0, \quad \Rightarrow \quad \mathbf{w} = \frac{1}{2}\sum_{i=1}^{n}\alpha_i y_i x_i, \\ \nabla_{\xi}L = C - \alpha_i - \lambda_i = 0, \quad \Rightarrow \quad C = \alpha_i + \lambda_i, \ \forall i. \end{aligned}$$

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

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

According to the complementary condition, we also have

- $y_i\mathbf{w}^\top x_i > 1 \quad \Rightarrow \quad \xi_i = 0 \quad \Rightarrow \quad \alpha_i = 0,$
- $y_i\mathbf{w}^\top x_i < 1 \quad \Rightarrow \quad \xi_i > 0 \quad \Rightarrow \quad \lambda_i = 0 \quad \Rightarrow \quad \alpha_i = C \ (\lambda_i = C - \alpha_i),$
- $y_i\mathbf{w}^\top x_i = 1 \quad \Rightarrow \quad \xi_i = 0 \quad \Rightarrow \quad \alpha_i \geq 0 \quad \text{and} \quad \lambda_i \geq 0 \quad \Rightarrow \quad \alpha_i \leq C \ (\lambda_i = C - \alpha_i),$

implying $0 \le \alpha_i \le C$, $\forall i$. (1 point)
Thus, the dual problem is

$$\max_{\alpha} \; -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i,$$
$$\text{s.t. } 0 \le \alpha_i \le C, \; \forall i. \quad \text{(1 point)}$$

(e) In order to handle non-linear classification, we consider to extend the dual problem in (d) based on an arbitrary kernel function $k(\cdot,\cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle$, in which $\phi(\cdot)$ maps the original feature vector into a high-dimensional space. (1 point)

$$\max_{\alpha} \; -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) + \sum_{i=1}^{n} \alpha_i,$$
$$\text{s.t. } 0 \le \alpha_i \le C, \; \forall i. \quad \text{(1 point)}$$

7. (13 points) *Neural network and backpropagation.*
Figure 4 shows a 2-layer, feed-forward neural network with two hidden-layer nodes and one output node. x_1 and x_2 are the two inputs. For the following questions, assume the learning rate η in gradient descent is fixed by $\eta = 0.1$. Each node also has a bias input value of +1. Assume there is a sigmoid activation function at the hidden layer nodes and at the output layer node. A sigmoid activation function takes the form: $g(z) = \frac{1}{1+e^{-z}}$, where $z = \sum_{j=1}^{d} w_j x_j$ and w_j is the jth incoming weight to a node, x_j is the jth incoming input value, and d is the number of incoming edges to the node.
Note: please round your results to 3 decimal places.

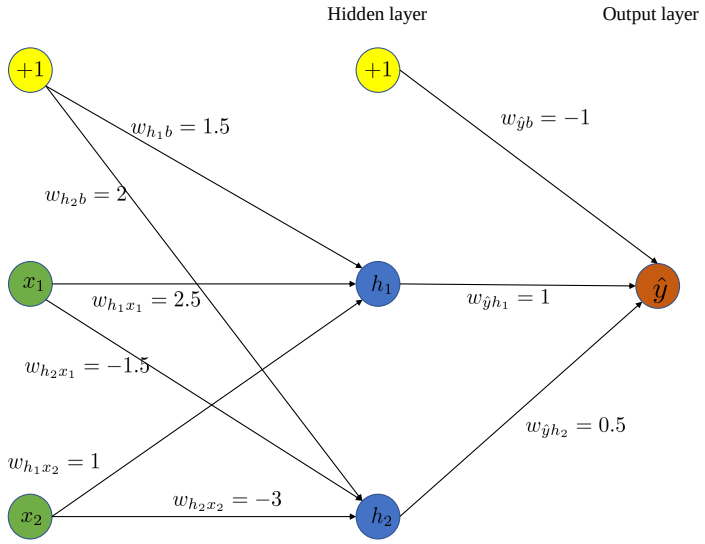


Figure 4: The neural network architecture.

(a) [5pts] Calculate the output values at nodes h_1, h_2 and $\hat{y}$ of this network for input $\{x_1 = 0, x_2 = 1\}$. Show all steps in your calculation. ($e \approx 2.7813$, $e^{-0.0586} \approx 0.9431$, $e^{-2.5} \approx 0.0821$.)

(b) [8pts] Compute one step of the backpropagation algorithm for a given example with input $\{x_1 = 0, x_2 = 1\}$ and target output $y = 1$. The network output is the real-valued output of the sigmoid function, so the error on the given example is defined as $E = \frac{1}{2}(y - \hat{y})^2$ where $\hat{y}$ is the real-valued network output of that example at the output node, and y is the integer-valued target output for that example. You are asked to compute the updated weights for the hidden layer h_1 and the output layer (note that there are six updated weights in total (i.e., the three incoming weights to node h_1 and the three incoming weights to node $\hat{y}$)) by performing ONE step of gradient descent. Show all steps in your calculation.

★ **Solution:**

(a) Let $g(a) = \frac{1}{1+e^{-a}}$
Output of node h_1 is $g(1 \times 1.5 + 0 \times 2.5 + 1 \times 1) = g(2.5) = 0.9241 \approx 0.924$. (2 points)
Output of node h_2 is $g(1 \times 2 + 0 \times (-1.5) + 1 \times (-3)) = g(-1) = 0.2689 \approx 0.269$. (2 points)
Output of node $\hat{y}$ is $g(1 \times (-1) + 0.9241 \times 1 + 0.2689 \times 0.5) = g(0.0586) = 0.5146 \approx 0.515$. (1 point)

(b)

$$\frac{\partial E}{\partial z_{\hat{y}}} = (g(z_{\hat{y}}) - y)\, g(z_{\hat{y}})\, (1 - g(z_{\hat{y}})) = (0.5146 - 1) \times 0.5146 \times (1 - 0.5146) = -0.1212.$$

(2 points)

$$\begin{aligned}
w_{\hat{y}h_1} &= 1 - (0.1)(-0.1212)(0.9241) = 1.0112 \approx 1.011, && \text{(1 point)} \\
w_{\hat{y}h_2} &= 0.5 - (0.1)(-0.1212)(0.2689) = 0.5033 \approx 0.503, && \text{(1 point)} \\
w_{\hat{y}b} &= -1 - (0.1)(-0.1212)(1) = -0.9879 \approx -0.988, && \text{(1 point)} \\
w_{h_1 x_1} &= 2.5 - (0.1)(-0.1212)(1)(0.9241)(1 - 0.9241)(0) = 2.500, && \text{(1 point)} \\
w_{h_1 x_2} &= 1 - (0.1)(-0.1212)(1)(0.9241)(1 - 0.9241)(1) = 1.0009 \approx 1.001, && \text{(1 point)} \\
w_{h_1 b} &= 1.5 - (0.1)(-0.1212)(1)(0.9241)(1 - 0.9241)(1) = 1.5009 \approx 1.501. && \text{(1 point)}
\end{aligned}$$

8. (13 points) *Convex sets and convex functions.*
In this problem, you should first write down whether the set or the function is convex, concave or neither, then you should either prove the set or the function is convex or provide an example to show that it's not convex. Correctly guessing whether the set is convex, concave or neither without proof will get 0 point.
Note: here we use $\mathbb{S}^n$ to denote the set of symmetric matrices in $\mathbb{R}^{n\times n}$, and $\mathbb{S}^n_+$ to denote the set of positive semi-definite symmetric matrices in $\mathbb{R}^{n\times n}$.

(a) [3pts] Determine the convexity of set $\mathcal{C}$:

$$\mathcal{C} = \{\mathbf{A} \in \mathbb{S}^n | \lambda_{\min}(\mathbf{A}) \geq 2\}, \tag{11}$$

where $\lambda_{\min}(\mathbf{A})$ refers to the minimum eigenvalue of $\mathbf{A}$, i.e.,

$$\lambda_{\min}(\mathbf{A}) = \min_{\|u\|_2=1} u^\top \mathbf{A} u. \tag{12}$$

(b) [4pts] Let $\mathcal{H}(w)$ denote the hyperplane with normal direction $w \in \mathbb{R}^d$, that is

$$\mathcal{H}(w) = \{x \in \mathbb{R}^d | x^\top w = 0\}. \tag{13}$$

Let $P : \mathbb{R}^d \to \mathbb{R}^d$ be given by

$$P(x) = \arg\min_{y\in\mathcal{H}(w)} \|y - x\|_2. \tag{14}$$

Determine the convexity of set $\mathcal{C}$:

$$\mathcal{C} = \{P(x) | x \in \mathcal{B}\}, \tag{15}$$

where $\mathcal{B} = \{x \in \mathbb{R}^d | \ \|x\|_2 \leq 1\}$.

(c) [6pts] Given a set of labeled data $S_\ell = \{(x_i, y_i)\}_{i=1}^n$, the ℓ_1-regularized support vector machines (SVM) considers the following unconstraint optimization problem:

$$\min_{\mathbf{w},w_0} \sum_{x_i\in S_\ell} \big(1 - y_i(\mathbf{w}^\top x_i + w_0)\big)_+ + \gamma\|\mathbf{w}\|_1, \tag{16}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the vector of model parameters, $w_0 \in \mathbb{R}$ is the bias, and $(\cdot)_+ = \max\{0, \cdot\}$. Suppose that additional unlabeled data $S_u = \{x_i\}_{i=1}^m$ is presented, and now we enable to extend (16) for handling semi-supervised learning by

$$\min_{\mathbf{w},w_0} \sum_{x_i\in S_\ell} \big(1 - y_i(\mathbf{w}^\top x_i + w_0)\big)_+ + \gamma\|\mathbf{w}\|_1 + \lambda \mathbf{w}^\top \mathbf{X}^\top \mathbf{L}\mathbf{X}\mathbf{w}, \tag{17}$$

where $\gamma, \lambda > 0$ are regularization parameters. In (17), the term $\mathbf{w}^\top\mathbf{X}^\top\mathbf{L}\mathbf{X}\mathbf{w}$ is *manifold regularization*, in which $\mathbf{X} \in \mathbb{R}^{(n+m)\times d}$ is the data matrix composed of both labeled and unlabeled data, and $\mathbf{L} \in \mathbb{S}^{(n+m)}$ denotes the normalized Laplacian matrix. Based on the fact that $\mathbf{X}^\top\mathbf{L}\mathbf{X} \in \mathbb{S}^d_+$, please determine the convexity of the objective function of (17).

★ **Solution**:

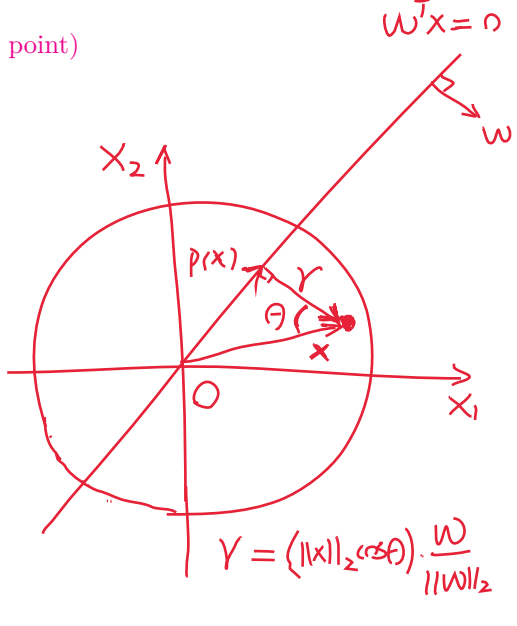
(a) For $\forall \mathbf{X}, \mathbf{Y} \in \mathcal{C}$ and $\forall \theta \in [0, 1]$, we have

$$\begin{aligned}
&\lambda_{\min}(\theta\mathbf{X} + (1-\mathbf{X})\mathbf{Y}) \\
=&\min_{u:\|u\|_2=1} u^\top(\theta\mathbf{X} + (1-\theta)\mathbf{Y})u \\
=&\theta \min_{u:\|u\|_2=1} u^\top\mathbf{X}u + (1-\theta)\min_{u:\|u\|_2=1} u^\top\mathbf{Y}u \quad \text{(1 point)} \\
\geq&2\theta + 2(1-\theta) \quad \text{(1 point)} \\
=&2.
\end{aligned}$$

So it's convex. (1 point)

(b) We can find that $\mathcal{C}$ is the projection set of $\mathcal{B}$. For $x \in \mathcal{B}$,

$$\begin{aligned}
P(x) &= x - (\|x\|_2 \cos\theta)\frac{w}{\|w\|_2} \\
&= x - (\|x\|_2 \frac{w^\top x}{\|w\|_2\|x\|_2})\frac{w}{\|w\|_2} \\
&= x - \frac{x^\top w}{w^\top w}w, \quad \text{(2 points)}
\end{aligned}$$



where θ is the acute angle between x and w.
Obviously, $P(x)$ is an affine transform of x. So $\mathcal{C}$ is an affine transform of $\mathcal{B}$. Since $\mathcal{B}$ is a convex set, $\mathcal{C}$ is a convex set. (2 points)
(Do not need to figure out the detailed formula of $P(x)$. Figuring out the projection is the affine transformation can get the full points.)

(c) Let $f(\mathbf{w})$ denote the objective function of (17), and we rewrite it as follows:

$$f(\mathbf{w}) = f_1(\mathbf{w}) + f_2(\mathbf{w}) + f_3(\mathbf{w}), \quad \mathrm{dom} f = \mathbb{R}^d,$$

where

$$\begin{aligned} f_1(\mathbf{w}) &= \sum_{x_i \in S_\ell} \left(1 - y_i(\mathbf{w}^\top x_i + w_0)\right)_+, \quad \mathrm{dom} f_1 = \mathbb{R}^d, \\ f_2(\mathbf{w}) &= \gamma \|\mathbf{w}\|_1, \quad \mathrm{dom} f_2 = \mathbb{R}^d, \\ f_3(\mathbf{w}) &= \lambda \mathbf{w}^\top \mathbf{X}^\top \mathbf{L} \mathbf{X} \mathbf{w}, \quad \mathrm{dom} f_3 = \mathbb{R}^d. \end{aligned}$$

- Hinge loss $(1 - y_i(\mathbf{w}^\top x_i + w_0))_+ = \max(0, 1 - y_i(\mathbf{w}^\top x_i + w_0))$ is convex because
  - any affine function is convex, e.g., 0 and $1 - y_i(\mathbf{w}^\top x_i + w_0)$;
  - the maximum of two convex functions is convex.

  Therefore, $f_1(\mathbf{w})$ is convex as summation preserves convexity. (2 points)
- The second component $f_2(\mathbf{w})$ is convex, because the ℓ_1 norm is convex, and nonnegative scaling preserves convexity. (1 point)
- The Hessian matrix of $f_3(\mathbf{w})$ w.r.t. $\mathbf{w}$ satisfies

  $$\nabla^2 f_3(\mathbf{w}) = \lambda \mathbf{X}^\top \mathbf{L} \mathbf{X} \succeq_{\mathbb{S}_+^d} 0,$$

  therefore, $f_3(\mathbf{w})$ is also convex. (2 points)

In conclusion, $f(\mathbf{w})$ is convex as it is a summation of three convex functions. (1 point)