Theoretical Assignment

Linear Algebra

1) A symmetric matrix A over $\mathbb{R}$ is called positive semidefinite (PSD) if for every vector v, $v^T A v \geq 0$.

a) Denote A is a real symmetric matrix. We will show that the following are equivalent:

(i) A is PSD.

(ii) A can be written as $A = XX^T$.

(iii) All $A's$ eigenvalues are non-negative.

Recall that a real symmetric matrix A can be decomposed as $A = QDQ^T$, where Q is an orthogonal matrix, whose columns are eigenvectors of A and D is a diagonal matrix with eigenvalues of A as its diagonal elements. We can prove that equivalence by showing the following argument: $(ii) \to (i) \to (iii) \to (ii)$.

$(ii) \to (i)$ Denote A can be written as $A = XX^T$ and let $v \in \mathbb{R}^n$. Then:

$$v^T A v = v^T X X^T v = (X^T v)^T (X^T v) =_{X^T v =: u \in \mathbb{R}^n} u^T u = \sum_{i=1}^{n} u_i^2 \geq 0 \blacksquare$$

$(i) \to (iii)$ Let $\lambda_1, \dots, \lambda_n$ be all the eigenvalues of A and let $b_1, \dots, b_n$ be its eigenvectors correspondingly. So, $Ab_j = \lambda_j b_j$ for every $j = 1, \dots, n$. Therefore, for every $j = 1, \dots, n$ by the property of (i):

$$0 \leq_{(i)} b_j^T A b_j = b_j^T \lambda_j b_j = \lambda_j b_j^T b_j$$

Since $b_j \neq 0$ as eigenvector, $b_j^T b_j = \sum_{k=1}^{n} \left(b_j^{(k)}\right)^2 > 0$ and therefore by dividing it we get:

$$\lambda_j \geq 0 \blacksquare$$

$(iii) \to (ii)$ Since A is symmetric matrix, by the recall we can write A as $A = QDQ^T$ as described above. Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of A as appear in the diagonal of D. From (iii) we know that $\lambda_i \geq 0$ for every $i = 1, \dots, n$. So, let's define $\sqrt{D} = \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_n} \end{pmatrix}$.

Therefore, $\sqrt{D}^2 = D$ and $\sqrt{D}^T = \sqrt{D}$. Hence, we can write A as $QDQ^T = Q\sqrt{D}\sqrt{D}Q^T = Q\sqrt{D}\sqrt{D}^T Q^T = \left(Q\sqrt{D}\right)\left(Q\sqrt{D}\right)^T$. Consider $X = Q\sqrt{D}$ and we've done. $\blacksquare$

(b) Let $\alpha, \beta \geq 0$ and PSD matrices $A, B \in \mathbb{R}^{n \times n}$, and let $v \in \mathbb{R}^n$. Then:

$$v^T(\alpha A + \beta B)v = (v^T \alpha A + v^T \beta B)v = (\alpha v^T A + \beta v^T B)v = \alpha v^T A v + \beta v^T B v \geq_* 0$$

(*) $v^T A v \geq 0, v^T B v \geq 0$ because A, B are PSD and $\alpha, \beta \geq 0$.

Therefore, $\alpha A + \beta B$ is also PSD matrix. It does not mean that all the set of PSD matrices is a vector space over the real numbers, because the coefficients α, β must be non-negative. For negative number we might get out of this set: Suppose $\alpha = -1$ and $A = I$. Indeed, I is PDS since $\forall v \in \mathbb{R} : v^T I v = v^T v \geq 0$. However, $\forall v \neq 0 : v^T(-1)Iv = -v^T I v = -v^T v < 0$. Therefore, $\alpha A = -I$ is not a PSD.

Calculus and Probability

1) Let $X_1, \dots, X_n$ be IID $U([0,1])$ continuous variables, and let $Y = \max(X_1, \dots, X_n)$.

(a) Notice that the max function Y is also continues function of $X_1, \dots, X_n$. Also notice that Y well defined only for y in $[0,1]$. Computing the PDF of Y – the function f_Y:

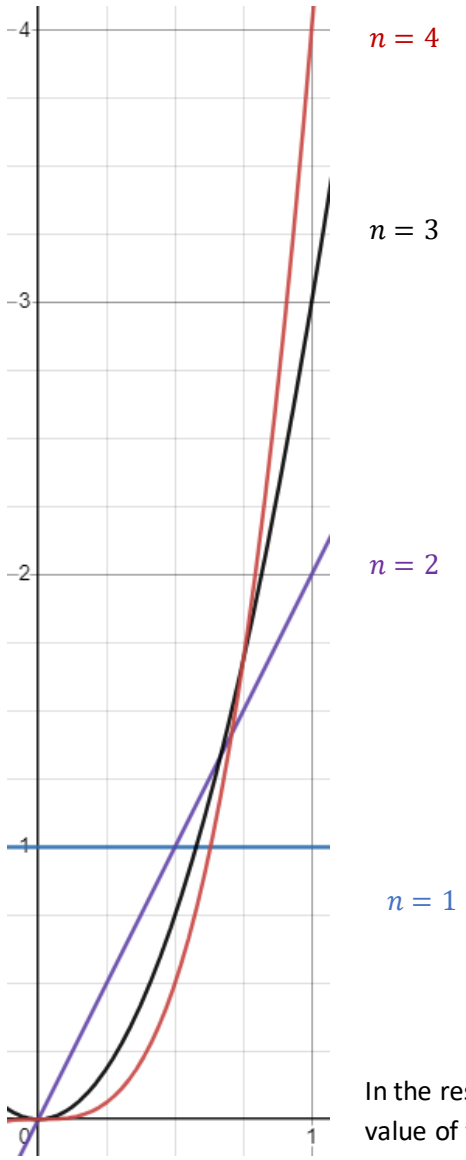
$$\forall y \in [0,1] \ : \ \mathbb{P}(Y \leq y) = F_Y(y) = \int_0^y f_Y(x)dx$$

On the other hand, for every $y \in [0,1]$:

$$\mathbb{P}(Y \leq y) = \mathbb{P}(\max(X_1, \dots, X_n) \leq y)$$
$$= \mathbb{P}(X_i \leq y \ for \ every \ i = 1, \dots, n) =_{IID} \left(\mathbb{P}(X_1 \leq y)\right)^n =_{uniform} y^n$$

Therefore:

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \begin{cases} \frac{d}{dy}y^n & y \in [0,1] \\ 0 & otherwise \end{cases} = \begin{cases} n \cdot y^{n-1} & y \in [0,1] \\ 0 & otherwise \end{cases}$$



In the rest of the real line the value of the function is 0

Computing $\mathbb{E}[Y]$:

$$\mathbb{E}[Y] = \int_0^1 y f_Y(y)dy = \int_0^1 ny^n dy = \left[n \cdot \frac{y^{n+1}}{n+1}\right]_0^1 = \frac{n}{n+1}$$

As n grows large, $\mathbb{E}[Y]$ is increasing and getting close to 1.

Computing $Var[Y]$:

$$Var[Y] = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2$$

- Computing f_{Y^2} as we did before:

$$\forall y \in [0,1] \ : \ \mathbb{P}(Y^2 \leq y) = F_{Y^2}(y) = \int_0^y f_{Y^2}(x)dx$$

$$\begin{aligned}\mathbb{P}(Y^2 \leq y) &= \mathbb{P}(\max(X_1, \dots, X_n)^2 \leq y) \\ &= \mathbb{P}\left(X_i^2 \leq y \ for \ every \ i = 1, \dots, n\right) =_{IID} \left(\mathbb{P}(X_1^2 \leq y)\right)^n \\ &= \left(\mathbb{P}\left(X_1 \leq \sqrt{y}\right)\right)^n =_{uniform} \sqrt{y}^n = y^{\frac{n}{2}}\end{aligned}$$

$$\Rightarrow \forall y \in [0,1): \ f_{Y^2}(y) = \frac{d}{dy} F_{Y^2}(y) = \frac{d}{dy} y^{\frac{n}{2}} = \frac{n}{2} \cdot y^{\frac{n}{2}-1}$$

$$\Rightarrow \mathbb{E}[Y^2] = \int_0^1 y f_{Y^2}(y)dy = \int_0^1 \frac{n}{2} \cdot y^{\frac{n}{2}} dy = \left[\frac{n}{2} \cdot \frac{y^{\frac{n}{2}+1}}{\frac{n}{2}+1}\right]_0^1 = \frac{\frac{n}{2}}{\frac{n}{2}+1}$$

Therefore,

$$\begin{aligned}Var(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 &= \frac{\frac{n}{2}}{\frac{n}{2}+1} - \frac{n^2}{(n+1)^2} = n \cdot \frac{(n+1)^2 - n(n+2)}{(n+2)(n+1)^2} \\ &= \frac{n}{(n+2)(n+1)^2}\end{aligned}$$

As n grows large, $Var(Y)$ is decreasing to 0 like $O\left(\frac{1}{n^2}\right)$ and $\mathbb{E}[Y]$ is increasing to 1.

Optimal Classifiers and Decision Rules

1)

(a) Let X and Y be random variables where Y can take values in $\mathcal{Y} = \{1, \dots, L\}$. Let ℓ_{0-1} be the 0-1 loss function defined in class.

Let's find h which gives the following minimum: $h = \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}\big[\ell_{0-1}\big(Y, f(X)\big)\big]$.

For every $f: \mathcal{X} \to \mathcal{Y}$:

$$\mathbb{E}\big[\ell_{0-1}\big(Y, f(X)\big)\big] =_{by\ \ell_{0-1}\ def} \sum_{i\in\mathcal{Y}, x\in\mathcal{X}} \mathbb{P}(Y = i\,, X = x)\,\Delta_{0-1}\big(i, f(x)\big)$$

For given $x \in \mathcal{X}$ let's find the optimal value $\hat{y} = h(x)$. The relevant part that depends on x after ignoring the multiplication constant $\mathbb{P}(X = x) \geq 0$:

$$\sum_{\substack{i\in\mathcal{Y}\\ i\neq h(x)}} \mathbb{P}(Y = i \mid X = x) = \begin{cases} \sum_{\substack{i\in\mathcal{Y}\\ i\neq 1}} \mathbb{P}(Y = i \mid X = x) & h(x) = 1 \\ \dots & \\ \sum_{\substack{i\in\mathcal{Y}\\ i\neq L}} \mathbb{P}(Y = i \mid X = x) & h(x) = L \end{cases}$$

Therefore, the minimal value is gotten where:

$$h(x) = \arg\max_{i\in\mathcal{Y}} \mathbb{P}(Y = i \mid X = x) \ \blacksquare$$

(b) Let X and Y be random variables where Y can take values in $\mathcal{Y} = \{0,1\}$. Let Δ be the following symmetric loss function:

$$\Delta(y, \hat{y}) = \begin{cases} 0 & y = \hat{y} \\ a & y = 0, \hat{y} = 1 \\ b & y = 1, \hat{y} = 0 \end{cases}$$

Where $a, b \in (0,1]$.

Let's find h which gives the following minimum: $h = \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}\big[\Delta\big(Y, f(X)\big)\big]$.

For every $f: \mathcal{X} \to \mathcal{Y}$:

$$\mathbb{E}\big[\Delta\big(Y, f(X)\big)\big] = \sum_{x\in\mathcal{X}} \Delta(0, f(x)) \cdot \mathbb{P}(Y = 0, X = x) + \Delta(1, f(x)) \cdot \mathbb{P}(Y = 1, X = x)$$

For given $x \in \mathcal{X}$ let's find the optimal value $\hat{y} = h(x)$. The relevant part that depends on x after ignoring the multiplication constant $\mathbb{P}(X = x) \geq 0$:

$$\Delta(0, h(x)) \cdot \mathbb{P}(Y = 0 \mid X = x) + \Delta(1, h(x)) \cdot \mathbb{P}(Y = 1 \mid X = x)$$
$$= \begin{cases} b \cdot \mathbb{P}(Y = 1 \mid X = x) & h(x) = 0 \\ a \cdot \mathbb{P}(Y = 0 \mid X = x) & h(x) = 1 \end{cases}$$

Therefore, the optimal value for every $x \in \mathcal{X}$ is:

$$h(x) = \begin{cases} 0 & b \cdot \mathbb{P}(Y = 1 \mid X = x) \leq a \cdot \mathbb{P}(Y = 0 \mid X = x) \\ 1 & else \end{cases}$$

2) Let X and Y be random variables where X can take values in some set $\mathcal{X}$ and Y can take values in $\mathcal{Y} = \{0, 1\}$ (i.e., binary label space). We wish to find a predictor $h : \mathcal{X} \to [0, 1]$ which minimizes $\mathbb{E}[\Delta_{\log}(Y, h(X))]$, where $\Delta_{\log}$ is the following loss function known as the log-loss:

$$\Delta_{\log}(y, \hat{y}) = -y log(\hat{y}) - (1 - y)\log(1 - \hat{y})$$

$$\mathbb{E}\big[\Delta_{\log}\big(Y, h(X)\big)\big]$$
$$= \sum_{x\in\mathcal{X}} \Delta_{\log}(0, h(x)) \cdot \mathbb{P}(Y = 0, X = x) + \Delta_{\log}(1, h(x))$$
$$\cdot\, \mathbb{P}(Y = 1, X = x)$$

For given $x \in \mathcal{X}$, let's find the optimal value $\hat{y} = h(x)$. The relevant part that depends on x after ignoring the multiplication constant $\mathbb{P}(X = x) \geq 0$:

$$g\big(h(x)\big) = \Delta_{\log}\big(0, h(x)\big) \cdot \mathbb{P}(Y = 0 \mid X = x) + \Delta_{\log}\big(1, h(x)\big) \cdot \mathbb{P}(Y = 1 \mid X = x)$$
$$= -\log\big(1 - h(x)\big) \cdot \mathbb{P}(Y = 0 \mid X = x) - \log\big(h(x)\big) \cdot \mathbb{P}(Y = 1 \mid X = x)$$

Where $g : (0,1) \to \mathbb{R}$.

We want to find the value of $\hat{y} = h(x) \in [0,1]$ which gives $g(\hat{y})$ to be minimal:

$$g(\hat{y}) = -\log(1 - \hat{y}) \cdot \mathbb{P}(Y = 0 \mid X = x) - \log(\hat{y}) \cdot \mathbb{P}(Y = 1 \mid X = x)$$

As $\hat{y} \to 0^+$, $g(\hat{y}) \to \infty$ and as $\hat{y} \to 1^-$ then $g(\hat{y}) \to \infty$. In every $[a, b] \subset (0,1)$ g is a continues and differentiable function with respect to $\hat{y}$, and we can find its global minimum as local minimum in $(0,1)$.

$$g'(\hat{y}) = \frac{1}{1 - \hat{y}} \mathbb{P}(Y = 0 \mid X = x) - \frac{1}{\hat{y}} \mathbb{P}(Y = 1 \mid X = x)$$
$$= \frac{\hat{y}\mathbb{P}(Y = 0 \mid X = x) - (1 - \hat{y})\, \mathbb{P}(Y = 1 \mid X = x)}{\hat{y}(1 - \hat{y})} = 0$$
$$\Rightarrow \hat{y}\mathbb{P}(Y = 0 \mid X = x) - (1 - \hat{y})\, \mathbb{P}(Y = 1 \mid X = x) = 0$$
$$\Rightarrow \hat{y}\big(\mathbb{P}(Y = 0 \mid X = x) + \mathbb{P}(Y = 1 \mid X = x)\big) - \mathbb{P}(Y = 1 \mid X = x) = 0$$
$$\text{We know that } \mathbb{P}(Y = 0 \mid X = x) + \mathbb{P}(Y = 1 \mid X = x) = 1$$
$$\Rightarrow \hat{y} = \mathbb{P}(Y = 1 \mid X = x)$$

This is the only local critical point, and therefore, this is the local and global minimum of g. We can also show that this is minimum by the second derivative or by the increasing and decreasing domains of g.

So, for every $x \in \mathcal{X}$:

$$h(x) = \mathbb{P}(Y = 1 \mid X = x)$$

3) Let X and Y be random variables taking values in $X = \mathbb{R}$ and $Y = \{0, 1\}$ respectively, and assume that given $Y = 0$, $X \sim N(\mu, \sigma_0^2)$, and similarly, given $Y = 1$, $X \sim N(\mu, \sigma_1^2)$, where $\sigma_0 \neq \sigma_1$. Also assume $\mathbb{P}[Y = 1] = p_1$. We find the optimal decision rule for this distribution and the zero-one loss.

We want to find $h : \mathbb{R} \to \{0,1\}$ which minimizes $\mathbb{E}\big[\ell_{0-1}\big(Y, h(X)\big)\big]$.

$$\mathbb{E}\big[\ell_{0-1}\big(Y, h(X)\big)\big]$$
$$= \sum_{x \in \mathcal{X}} \Delta_{0-1}\big(0, h(x)\big)\mathbb{P}(Y = 0, X = x) + \Delta_{0-1}\big(1, h(x)\big)\mathbb{P}(Y = 1, X = x)$$

For given $x \in \mathbb{R}$ The relevant part that depends on x after ignoring the multiplication constant $\mathbb{P}(X = x) \geq 0$:

$$\Delta_{0-1}\big(0, h(x)\big)\mathbb{P}(Y = 0 \mid X = x) + \Delta_{0-1}\big(1, h(x)\big)\mathbb{P}(Y = 1 \mid X = x)$$
$$= \begin{cases} \mathbb{P}(Y = 1 \mid X = x) & h(x) = 0 \\ \mathbb{P}(Y = 0 \mid X = x) & h(x) = 1 \end{cases}$$

As we saw in class, the decision rule h that minimizes this expression is:

$$h(x) = \begin{cases} 0 & \mathbb{P}(Y = 1 \mid X = x) < \mathbb{P}(Y = 0 \mid X = x) \\ 1 & otherwise \end{cases}$$

Therefore, we want to find a condition for $\mathbb{P}(Y = 1 \mid X = x) < \mathbb{P}(Y = 0 \mid X = x)$ using the terms of μ, σ_0, σ_1 and p_1.

By using Bayes theorem for continues variable and plugging in the given, we get:

$$h(x) = 0 \Leftrightarrow \mathbb{P}(Y = 1 \mid X = x) < \mathbb{P}(Y = 0 \mid X = x)$$
$$\Leftrightarrow f_X(x|Y = 1) \cdot \frac{\mathbb{P}(Y = 1)}{f_X(x)} < f_X(x|Y = 0) \cdot \frac{\mathbb{P}(Y = 0)}{f_X(x)}$$

$$\Leftrightarrow \frac{f_X(x|Y=1)}{f_X(x|Y=0)} < \frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \Leftrightarrow \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_0^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right)} < \frac{1-p_1}{p_1}$$

$$\Leftrightarrow \exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2} + \frac{(x-\mu)^2}{2\sigma_0^2}\right) < \frac{1-p_1}{p_1} \cdot \frac{\sqrt{2\pi\sigma_1^2}}{\sqrt{2\pi\sigma_0^2}}$$

$$\Leftrightarrow (x-\mu)^2\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right) < \log\left(\frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right|\right)$$

$$\Leftrightarrow \left(\frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2}\right)(x-\mu)^2 - \log\left(\frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right|\right) < 0$$

The solutions of the corresponded equation:

$$\left(\frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2}\right)(x-\mu)^2 - \log\left(\frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right|\right) = 0 \Leftrightarrow_{\sigma_0 \neq \sigma_1} (x-\mu)^2 = \frac{\log\left(\frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right|\right)}{\left(\frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2}\right)}$$

$$\Leftrightarrow (x-\mu)^2 = \frac{2\sigma_0^2\sigma_1^2 \log\left(\frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right|\right)}{\sigma_1^2 - \sigma_0^2}$$

Separate into 3 cases:

(1) If $\frac{2\sigma_0^2\sigma_1^2 \log\left(\frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right|\right)}{\sigma_1^2 - \sigma_0^2} < 0$ then there is no solution for the equation. Therefore, if $\sigma_1 > \sigma_0$ then there is no solution to the inequality. And if $\sigma_1 < \sigma_0$ every $x \in \mathbb{R}$ is a solution for the inequality.

Notice that $\frac{2\sigma_0^2\sigma_1^2 \log\left(\frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right|\right)}{\sigma_1^2 - \sigma_0^2} < 0 \Leftrightarrow \frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right| < 1$ and $\sigma_1 > \sigma_0$ **or** $\frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right| > 1$ and $\sigma_1 < \sigma_0 \Leftrightarrow (1-p_1) \cdot \left|\frac{\sigma_1}{\sigma_0}\right| - p_1 < 0$ and $\sigma_1 > \sigma_0$ **or** $(1-p_1) \cdot \left|\frac{\sigma_1}{\sigma_0}\right| - p_1 > 0$ and $\sigma_1 < \sigma_0 \Leftrightarrow \frac{\left|\frac{\sigma_1}{\sigma_0}\right|}{1+\left|\frac{\sigma_1}{\sigma_0}\right|} < p_1$ and $\sigma_1 > \sigma_0$ **or** $\frac{\left|\frac{\sigma_1}{\sigma_0}\right|}{1+\left|\frac{\sigma_1}{\sigma_0}\right|} > p_1$ and $\sigma_1 < \sigma_0$

Therefore,

$$h(x) = \begin{cases} 1 & \frac{\left|\frac{\sigma_1}{\sigma_0}\right|}{1+\left|\frac{\sigma_1}{\sigma_0}\right|} < p_1 \text{ and } \sigma_1 > \sigma_0 \\ 0 & \frac{\left|\frac{\sigma_1}{\sigma_0}\right|}{1+\left|\frac{\sigma_1}{\sigma_0}\right|} > p_1 \text{ and } \sigma_1 < \sigma_0 \end{cases}$$

(2) If $\frac{2\sigma_0^2\sigma_1^2 \log\left(\frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right|\right)}{\sigma_1^2 - \sigma_0^2} = 0$ then $x = \mu$ is the only solution for the equation. Therefore, if $\sigma_1 > \sigma_0$ then there is no solution to the inequality. And if $\sigma_1 < \sigma_0$ every $\mu \neq x \in \mathbb{R}$ is a solution for the inequality.

Notice that $\frac{2\sigma_0^2\sigma_1^2 \log\left(\frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right|\right)}{\sigma_1^2 - \sigma_0^2} = 0 \Leftrightarrow \frac{1-p_1}{p_1} \cdot \left|\frac{\sigma_1}{\sigma_0}\right| = 1 \Leftrightarrow \frac{\left|\frac{\sigma_1}{\sigma_0}\right|}{1+\left|\frac{\sigma_1}{\sigma_0}\right|} = p_1$. In this scenario when $\frac{\left|\frac{\sigma_1}{\sigma_0}\right|}{1+\left|\frac{\sigma_1}{\sigma_0}\right|} = p_1$ we get:

$$h(x) = \begin{cases} 1 & \sigma_1 > \sigma_0 \\ 0 & \sigma_1 < \sigma_0 \text{ and } x \neq \mu \\ 1 & \sigma_1 < \sigma_0 \text{ and } x = \mu \end{cases}$$

(3) For the other cases, we have 2 different solutions for the equation:

$$x_{1,2} = \pm\sqrt{\frac{2\sigma_0^2\sigma_1^2\log\left(\frac{1-p_1}{p_1}\cdot\left|\frac{\sigma_1}{\sigma_0}\right|\right)}{\sigma_1^2-\sigma_0^2}} + \mu = \pm\sqrt{2}|\sigma_0||\sigma_1|\sqrt{\frac{\log\left(\frac{1-p_1}{p_1}\cdot\left|\frac{\sigma_1}{\sigma_0}\right|\right)}{\sigma_1^2-\sigma_0^2}} + \mu$$

Consider $x_1 < x_2$.

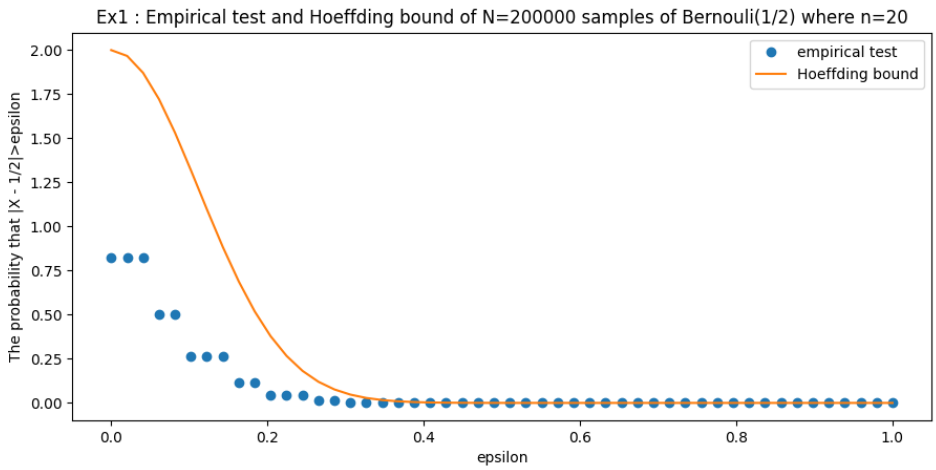
If $\sigma_1 > \sigma_0$ then the solution of the inequality is $x_1 < x < x_2$ and otherwise $x < x_1 \; or \; x > x_2$.

So, the function in this case will be:

$$h(x) = \begin{cases} 0 & \sigma_1 > \sigma_0 \;\; and \; x_1 < x < x_2 \\ 1 & \sigma_1 > \sigma_0 \; otherwise \\ 0 & \sigma_1 < \sigma_0 \; \boldsymbol{and} \; x < x_1 \; or \; x > x_2 \\ 1 & \sigma_1 < \sigma_0 \; otherwise \end{cases}$$

Programming Assignment

(1) Visualizing the Hoeffding bound. Output plot:



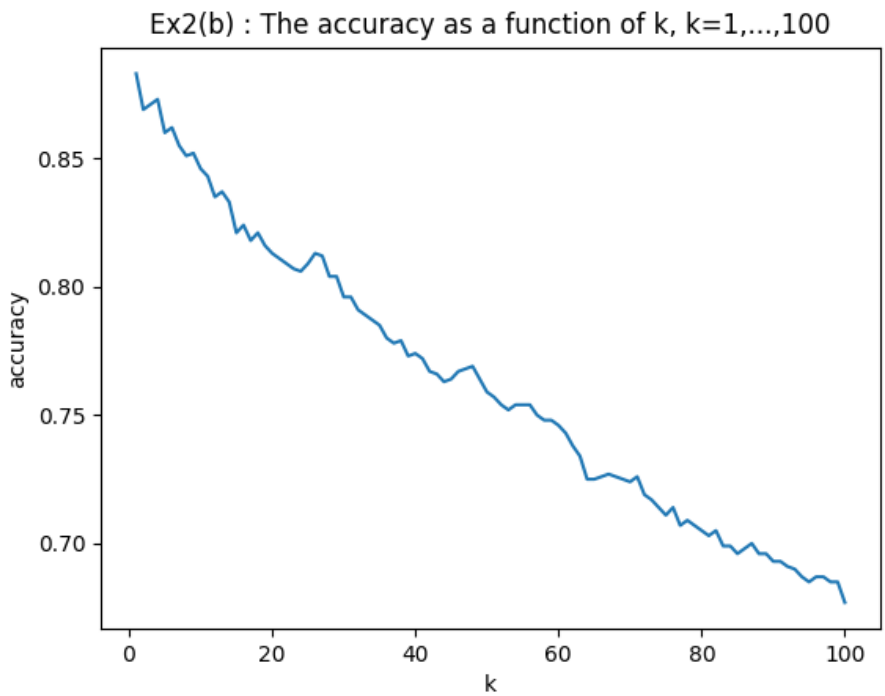
(2) KNN algorithm

(a) The code is attached in separated assignment box.

(b) The accuracy of the prediction is 0.846 (84.6%). I expect from completely random predictor to give us accuracy of 10%:

$$X \sim U(\{0,1,\dots,9\}) \Rightarrow \mathbb{P}(success) = \frac{1}{10}$$

(c) The accuracy as a function of k:

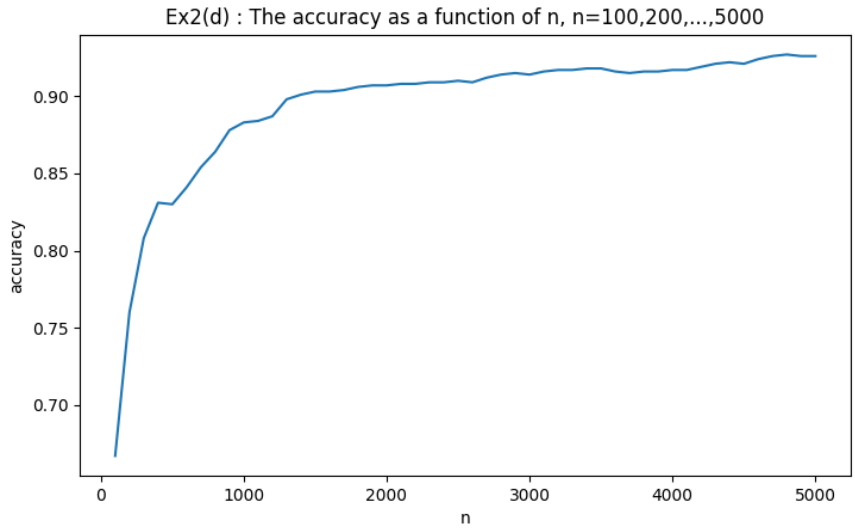


As we can see, the accuracy is tending to decrease when k grows large and the best k which gives us the best accuracy is k=1 (~87%).

k=1 means that the k_NN algorithm predicts the label of the closest trained image. With no other knowledge and with the Euclidean L2 matric, this is obviously the best we can do, as we want find a perfect match between the test image and the train images. k=100 means that the algorithm takes the common label of the 100 closest trained images to the test image. With train data set in size of 1000 images, we want

to assume that each digit has ~100 train images. However, it is not necessarily the reality. If one digit has much less than 100 images in the trained data set, there will be images that don't represented this digit in the top 100 closest images of this image (as test). Therefore, we might get an error for those test images. In addition, with a little bit knowledge about the shapes of digits, we can say that there are digits that closer to each other than others. For example, 1 and 7 or 5,6,8 and 9. Hence, for larger k, we might get in our top k closest images list more incorrect and resembles digits which leads to an error predict.

(d) The accuracy as a function of n:



As we can see, the accuracy is tending to increase as n grows large and the best n which gives us the best accuracy is n=5000 (~92%). We run the k_NN algorithm with k=1, and as we discussed previously, it means we choose the label of the closest image from the train data set to the test image. Obviously, as n grows large, we have more and more images on the data set, and therefore, we can find closer image to the test image.

Something interesting we can find in the plot is related to the accuracy growth rate. Initially, from n=100 to n=400 we have the biggest growth rate (from ~66% to ~83%), with linear approximate rate of 0.056. Then, from n=400 to n=1200 the growth rate becomes a bit smaller (from ~83% to ~89%), with linear approximate rate of 0.0075. Finally, the growth rate becomes much smaller until n=5000 (from ~89% to ~92%), linear approximate rate of 0.00078.

This behavior related to 2 parameters: (i) the size of the train data set; (ii) the simplicity of the data set objects. As I said before, as the train data set get wealthier, then we get a better accuracy. When n is small, we have more errors since we probably don't have enough examples for each digit in the train data set, and consequently we might do a bad predict. Nevertheless, there is a point such that we get enough examples for good performance (around 85%). We must remember that our image object is represented as 28*28 vector in grayscale color. Therefore, (*) by enough wealthy train data set, we can cover enough possibilities of potential test digits. So, adding more images to the train data set, would almost not improve the accuracy.

(*) By enough wealthy train data set, we can cover enough possibilities of potential test digits. For example, the digits 1 and 7 have enough points in data set that eliminate the closeness of 1 and 7. A test as shown in black can easily and correctly be predicted.