Intro to ML - HW1 - Theoretical part

Daniel Volkov, I.D: 330667494

June 11, 2024

Linear algebra

Note: I'll first solve (b), then use it for (a)'s solution.
General notation for this part:
The matrix $A \in \mathbb{R}^{n\times n}$ is symmetric, so from linear algebra, we conclude there exists an orthonormal basis $B = \{b_1, \dots, b_n\}$,such that:

$$[A]_B = D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \tag{1}$$

Where $\lambda_1 \dots \lambda_n$ are the eigenvalues of A. (Note that $b_1, \dots, b_n$ are eigen-vectors of A, with the corresponding eigen-values $\lambda_1, \dots, \lambda_n$)

Question (b)

$\Rightarrow$:
Suppose A is PSD. this means $\forall v \in \mathbb{R}^n : \ v^t Av \geq 0$.

This implies that $\forall i \in \{1, \dots, n\}$:

$$b_i^t Ab_i = \langle Ab_i, b_i\rangle = \langle \lambda_i b_i, b_i\rangle = \lambda_i \|b_i\| \geq 0$$

of course, $\forall v \in B : \ \|v\| > 0$ (a basis will not include the zero vector), and thus we conclude:

$$\forall i \in \{1, \dots, n\} : \ \lambda_i \geq 0$$

$\Leftarrow$:
Suppose all eigenvalues of A are non-negative.
Let $v \in \mathbb{R}^n$ be an arbitrary vector.
Denote $[v]_B = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$. This means: $v = \sum_{i=1}^n a_i b_i$.
This in turn means that:

$$v^t Av = \langle Av, v\rangle = \langle A \sum_{i=1}^n a_i b_i, \sum_{j=1}^n a_j b_j\rangle = \sum_{i=1}^n \sum_{j=1}^n \langle Aa_i b_i, a_j b_j\rangle =$$
$$= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \langle Ab_i, b_j\rangle = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \lambda_i \langle b_i, b_j\rangle$$

And since the basis B is orthonormal, $\langle b_i, b_j \rangle$ is 1 if $i = j$, and 0 otherwise. This means that:

$$v^t A v = \cdots = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \lambda_i \langle b_i, b_j \rangle = \sum_{i=1}^{n} a_i^2 \lambda_i \geq 0$$

And since $v \in \mathbb{R}^n$ was arbitrary, we are done.
$\blacksquare$

Question (a)

I'll use the same denotations here.
$\Rightarrow$:
Suppose A is PSD. from (b), all of it's eigen-values are non-negative.
Choose X such that $[X]_B = \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n} \end{pmatrix}$. It's easy to see that
$[X]_B = [X]_B^t$ and that $[X]_B [X]_B^t = [A]_B$.
From linear algebra we know that change of basis is an invertible linear function, so we can guarantee that such X exists.
We conclude:

$$[X]_B [X]_B^t = [XX^t]_B = [A]_B \implies XX^t = A$$

$\Leftarrow$:
Assume there exists a matrix $X \in \mathbb{R}^{n \times n}$ such that $A = XX^t$.
Let $v \in \mathbb{R}^n$ be an arbitrary vector.
We get:

$$v^t A v = \langle Av, v \rangle = \langle XX^t v, v \rangle = \langle X^t v, X^t v \rangle = \|X^t v\| \geq 0$$

Meaning that A is PSD.
$\blacksquare$

Question (c)

Let $v \in \mathbb{R}^n$ be an arbitrary vector.
We can easily see that:

$$v^t (\alpha A + \beta B) v = \alpha v^t A v + \beta v^t B v \geq 0$$

When the last inequality stands from the given that A and B are PSD.
This does **not** mean that the PSD matrices are a vector space over $\mathbb{R}$, because of negative scalars.
For example, we can easily see that the identity matrix I_n is PSD, but the matrix $I_n - 2I_n = -I_n$ is not.

Calculus and probability

Question 1

If we denote: $A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}$,and $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$
We can easily observe that:

$$y(x) = x^t A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

And so the partial derivatives are:

$$\begin{aligned} \frac{\partial y}{\partial x_i} &= \frac{\partial}{\partial x_i} (\sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} x_j x_k) = \frac{\partial}{\partial x_i} (\sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{k=1}^{n} a_{ki} x_i x_k) \\ &= \frac{\partial}{\partial x_i} (\sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{j=1}^{n} a_{ji} x_i x_j) = \frac{\partial}{\partial x_i} (x_i \cdot (\sum_{j=1}^{n} (a_{ij} + a_{ji}) \cdot x_j)) \\ &= \sum_{j=1}^{n} (a_{ij} + a_{ji}) \cdot x_j = \sum_{j=1}^{n} (A + A^t)_{ij} \cdot x_j = ((A + A^t) x)_i \end{aligned}$$

And so we conclude that (by the given definition):

$$\frac{\partial y}{\partial x} = (A + A^t) x$$

Question 2

As we know, the PDF of a random variable is determined by it's CDF. Here, the CDF is easy to compute:

$$\mathbb{P}(Y \leq a) = \mathbb{P}(\{X_1 \leq a\} \cap \dots \cap \{X_n \leq a\}) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq a)$$

We also know that $X_1, \dots, X_n$ uniformly distribute over $[0, 1]$ and thus:

$$\mathbb{P}(Y \leq a) = \prod_{i=1}^{n} \int_{-\infty}^{a} PDF_{X_i}(t) dt = \prod_{i=1}^{n} a = a^n$$

And so we calculate the PDF:

$$\mathbb{P}(Y \leq a) = \int_{-\infty}^{a} PDF_Y(t) dt = \int_{0}^{a} PDF_Y(t) dt = a^n$$

And we conclude (via fundamental theorem of calculus):

$$PDF_Y(t) = \begin{cases} n t^{n-1} & 0 \leq t \leq 1 \\ 0 & else \end{cases}$$

Now for the expected value:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} t \cdot PDF_Y(t)dt = \int_0^1 t \cdot nt^{n-1}dt = n\int_0^1 t^n dt = \frac{n}{n+1}$$

And the variance (calculate only over $[0,1]$ just as above):

$$Var(Y) = \int_0^1 t^2 \cdot PDF_Y(t)dt - \mathbb{E}[Y]^2 = n \cdot \int_0^1 t^{n+1} - (\frac{n}{n+1})^2$$
$$= \frac{n}{n+2} - (\frac{n}{n+1})^2 = \frac{n}{(n+2)(n+1)^2}$$

And as we can see:

$$\lim_{n\to\infty} \mathbb{E}[Y] = \lim_{n\to\infty} \frac{n}{n+1} = 1$$
$$\lim_{n\to\infty} Var(Y) = \lim_{n\to\infty} \frac{n}{(n+2)(n+1)^2} = 0$$

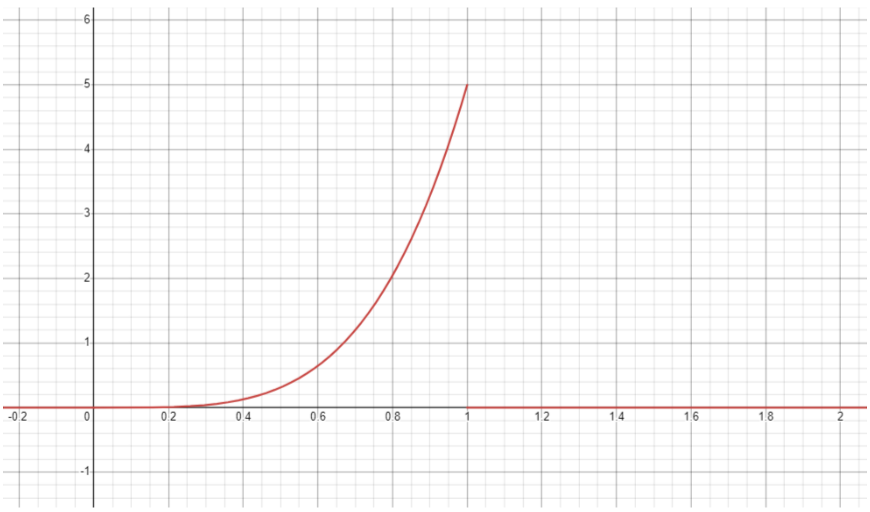

Figure 1: Graph of $PDF_Y(t)$ with $n = 5$

Optimal classifiers and decision rules

Question 1

(a)

Let's examine the expected loss for a given $\hat{x}$:

$$\mathbb{E}[\ell_{0-1}(Y, f(\hat{x}))] = \sum_{y=1}^{L} \mathbb{P}(X = \hat{x}, Y = y) \cdot \ell_{0-1}(y, \hat{x}) =$$
$$= \sum_{\substack{1 \leq y \leq L \\ y \neq f(\hat{x})}} \mathbb{P}(X = \hat{x}, Y = y) = 1 - \mathbb{P}(X = \hat{x}, Y = f(\hat{x}))$$

And so, if we want to minimize the expected loss, we should <u>maximize</u> the last probability. Meaning:

$$\arg\min_{f(\hat{x})} \mathbb{E}[\ell_{0-1}(Y, f(\hat{x}))] = \arg\max_{y} \mathbb{P}(X = \hat{x}, Y = y) = \arg\max_{y \in \{1,\dots,L\}} \mathbb{P}(Y = y | X = \hat{x}) \cdot \mathbb{P}(X = \hat{x}) =$$
$$= \arg\max_{y \in \{1,\dots,L\}} \mathbb{P}(Y = y | X = \hat{x})$$

Meaning we got the wanted result:

$$\forall \hat{x} \in \mathcal{X} : \; h(\hat{x}) = \arg\max_{y \in \{1,\dots,L\}} \mathbb{P}(Y = y | X = \hat{x})$$

■

(b)

Again let's inspect the expected loss for some $\hat{x} \in \mathcal{X}$:

$$\mathbb{E}[\Delta(Y, f(\hat{x}))] = \sum_{y=0}^{1} \mathbb{P}(Y = y, X = \hat{x}) \cdot \Delta(y, f(\hat{x})) =$$
$$= \mathbb{P}(Y = 0, X = \hat{x}) \cdot (0, f(\hat{x})) + \mathbb{P}(Y = 1, X = \hat{x}) \cdot \Delta(1, f(\hat{x})) =$$
$$= \begin{cases} \mathbb{P}(Y = 0, X = \hat{x}) \cdot a & f(\hat{x}) = 1 \\ \mathbb{P}(Y = 1, X = \hat{x}) \cdot b & f(\hat{x}) = 0 \end{cases}$$

Now denote $p = \mathbb{P}(Y = 0, X = \hat{x}) \implies 1 - p = \mathbb{P}(Y = 1, X = \hat{x})$
And so if we want to minimize the expected loss, we should decide:

$$\forall \hat{x} \in \mathcal{X} : \; h(\hat{x}) = \begin{cases} 1 & p \cdot a < (1-p) \cdot b \\ 0 & else \end{cases}$$

Question 2

Again let's look on some $\hat{x} \in \mathcal{X}$.

$$\mathbb{E}[\Delta(Y, f(\hat{x}))] = \sum_{y=0}^{1} \mathbb{P}(Y = y, X = \hat{x}) \cdot (-y \log(f(\hat{x})) - (1-y)\log(1 - f(\hat{x})) =$$
$$= \mathbb{P}(Y = 0, X = \hat{x}) \cdot (-\log(1 - f(\hat{x})) + \mathbb{P}(Y = 1, X = \hat{x}) \cdot (-\log(f(\hat{x})))$$

Now if we denote
$p = \mathbb{P}(Y = 1, X = \hat{x}) \implies 1 - p = \mathbb{P}(Y = 1, X = \hat{x})$, we get the function:

$$g(f(\hat{x})) = \mathbb{E}[\Delta(Y, f(\hat{x}))] = -((1-p)\log(1 - f(\hat{x})) + p\log(f(\hat{x})))$$

And remember that we want to minimize exactly $g(f(\hat{x}))$, and so we can approach this problem using standard calculus:

$$g'(x) = \frac{d}{dx}(-(1-p)\log(1-x) - p\log(x)) = \frac{1-p}{1-x} - \frac{p}{x} = \frac{x-p}{x-x^2}$$

And so the extrema point would be at $x = p = \mathbb{P}(Y = 1, X = \hat{x})$.
We can easily verify that this is a minimum extrema using the second derivative:

$$g''(x) = \frac{d}{dx}\left(\frac{1-p}{1-x} - \frac{p}{x}\right) = \frac{1-p}{(1-x)^2} + \frac{p}{x^2} \implies$$
$$g''(p) = \frac{1-p}{(1-p)^2} + \frac{p}{p^2} = \frac{1}{1-p} + \frac{1}{p} > 0$$

And so, in conclusion, we got that for $f(\hat{x}) = \mathbb{P}(Y = 1, X = \hat{x})$ we recieve minimum expected loss. Thus we conclude that the optimal classifier here would be:

$$\forall \hat{x} \in \mathcal{X}: \ h(\hat{x}) = \mathbb{P}(Y = 1, X = \hat{x})$$

Question 3

As we saw (in class and in Question 1 here), the optimal classifier for binary classification with the zero-one loss is given by:

$$h(x) = \arg\max_{y \in \{0,1\}} \mathbb{P}(Y = y | X = x) = \begin{cases} 0 & \mathbb{P}(Y = 0|X = x) > \mathbb{P}(Y = 1|X = x) \\ 1 & else \end{cases}$$

Now we can use Bayes' rule to check this condition:

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

And so:

$$\mathbb{P}(Y=0|X=x) > \mathbb{P}(Y=1|X=x) \iff$$

$$\frac{f_X(x|Y=0)\cdot f_X(x)}{\mathbb{P}(Y=0)} > \frac{f_X(x|Y=1)\cdot f_X(x)}{\mathbb{P}(Y=1)} \iff$$

$$\frac{f_X(x|Y=0)}{f_X(x|Y=1)} > \frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \iff$$

$$\frac{\frac{1}{\sigma_0\sqrt{2\pi}}\cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma_0})^2}}{\frac{1}{\sigma_1\sqrt{2\pi}}\cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma_1})^2}} > \frac{1-p_1}{p_1} \iff$$

$$\frac{\sigma_1}{\sigma_0}\cdot e^{\frac{(x-\mu)^2}{2\sigma_1^2}-\frac{(x-\mu)^2}{2\sigma_0^2}} > \frac{1-p_1}{p_1} \iff$$

$$\frac{(x-\mu)^2}{2\sigma_1^2} - \frac{(x-\mu)^2}{2\sigma_0^2} > \ln(\frac{(1-p_1)\cdot\sigma_0}{p_1\cdot\sigma_1})$$

Now if $\sigma_1 < \sigma_0$, we get that $\sigma_0^2 - \sigma_1^2 > 0$, which implies:

$$\mathbb{P}(Y=0|X=x) > \mathbb{P}(Y=1|X=x) \iff$$

$$(x-\mu)^2 > \frac{\ln(\frac{(1-p_1)\cdot\sigma_0}{p_1\cdot\sigma_1})\cdot 2\sigma_0^2\sigma_1^2}{\sigma_0^2-\sigma_1^2}$$

And on the other end, if $\sigma_1 > \sigma_0$, we get the same inequalities, with reversed sign.
Let's denote $a = \frac{\ln(\frac{(1-p_1)\cdot\sigma_0}{p_1\cdot\sigma_1})\cdot 2\sigma_0^2\sigma_1^2}{\sigma_0^2-\sigma_1^2}$.
Now, if $a \geq 0$, we can take the square root, and we get:

$$\mathbb{P}(Y=0|X=x) > \mathbb{P}(Y=1|X=x) \iff$$

$$|x-\mu| > \sqrt{a} \iff$$

$$x > \mu + \sqrt{a} \ \vee \ x < \mu - \sqrt{a}$$

On the other end, if $a < 0$, we want to classify eveything in the same class, depending on the sign of $\sigma_0 - \sigma_1$.

Thus, we got our optimal classifier(s), which I'll show in each case, when denoting $a = \frac{\ln(\frac{(1-p_1)\cdot\sigma_0}{p_1\cdot\sigma_1})\cdot 2\sigma_0^2\sigma_1^2}{\sigma_0^2-\sigma_1^2}$:

<u>If $\sigma_0 > \sigma_1$</u> :

$$\forall x \in \mathbb{R}: \; h(x) = \begin{cases} 1 & a \geq 0 \;\wedge\; x \in (\mu - \sqrt{a}, \mu + \sqrt{a}) \\ 0 & else \end{cases}$$

<u>If $\sigma_0 < \sigma_1$</u> :

$$\forall x \in \mathbb{R}: \; h(x) = \begin{cases} 0 & a \geq 0 \;\wedge\; x \in (\mu - \sqrt{a}, \mu + \sqrt{a}) \\ 1 & else \end{cases}$$