Linear Algebra

1.

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is called positive semidefinite (PSD) if for every vector $v \in \mathbb{R}^n$ we have $v^T Av \geq 0$.

a. Show that a symmetric matrix A is PSD if and only if it can be written as $A = XX^T$.

Solution:
($\Rightarrow$) We assume that A is PSD. Because A is a symmetric, it can be orthogonally diagonalized, Meaning $P^t AP = D$ where P is an orthogonal matrix, and $D = diag(\lambda_1, ..., \lambda_n)$. For any vector $v \in \mathbb{R}^n$ we define $y := P^t v$, then $Py = PP^t v = v$. And so

$$\langle v, Av\rangle = \langle Pv, APy\rangle = \langle y, P^t APy\rangle = \langle y, Dy\rangle$$

And so $\langle v, Av\rangle = \sum_{i=1}^n \lambda_i y_i^2$. From this we conclude that $\lambda_i \geq 0$ for every $i = 1, ..., n$. So we'll define $\sqrt{D} := diag(\sqrt{\lambda_1}, ..., \sqrt{\lambda_n})$ and define $X = P\sqrt{D}P^t$:

$$XX^t = P\sqrt{D}P^t(P^t)^t\sqrt{D}^t P^t = P\sqrt{D}\sqrt{D}P^t = PDP^t = A$$

So A can be written as $A = XX^t$ as required.
($\Leftarrow$) We assume that $A = XX^t$. Given some vector $v \in \mathbb{R}^n$:

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

Because for every $x \in \mathbb{R}^n, \langle x, x\rangle \geq 0$. Meaning A is PSD.
$\square$

b. Show that a symmetric matrix A is PSD if and only if all of its eigenvalues are non-negative.

Solution:
($\Rightarrow$) We assume that A is PSD. Given $Av = \lambda v$ for some eigenvector $v \neq 0$, then $0 \leq v^t Av = v^t \lambda v = \lambda v^t v$, and so $\lambda \geq 0$.
($\Leftarrow$) We assume all of A's eignevalues are non-negative. A is symmetric and so can be orthogonally diagonalized, meaning $PAP^t = D$ such that P is an orthogonal matrix, and $D = diag(\lambda_1, ..., \lambda_n)$. And so, for every vector $v \neq 0$ it follows that:

$$v^t Av = v^t P^t DPv = (Pv)^t D(Pv) \overset{(*)}{=} y^t Dy = \sum_{i=1}^n \lambda_i y_i^2 \geq 0$$

Meaning A is PSD.
$(*)$ We define $y := Pv$
$\square$

c. Show that for all $\alpha, \beta \geq 0$ and PSD matrices $A, B \in \mathbb{R}^{n\times n}$, the matrix $\alpha A + \beta B$ is also PSD. Does this mean that the set of all $n \times n$ PSD matrices is a vector space over $\mathbb{R}$?

Solution:
A and B are PSD, and so they're both symmetric, and so $(\alpha A+\beta B)^t = \alpha A^t+\beta B^t = \alpha A+\beta B$, so $\alpha A + \beta B$ is symmetric.
For every vector $v \in \mathbb{R}^n$:

$$v^t(\alpha A + \beta B)v = \alpha v^t Av + \beta v^t Bv$$

Because A and B are PSD, $\alpha v^t Av \geq 0$ and $\beta v^t Bv \geq 0$ and so:

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

Meaning $\alpha A + \beta B$ is PSD.

The set of all $n \times n$ PSD matrices is not a vector space over $\mathbb{R}$ because it's not closed under scalar multiplication. Take $A = I \in \mathbb{R}^2$. A is PSD but given $\alpha = -1$, the matrix $\alpha A = -I$ is not PSD, for $v = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$:

$$\begin{pmatrix} 1 & 0 \end{pmatrix} \cdot \alpha A \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = -1 \ngeq 0$$

$\square$

Calculus and Probability

1.

Matrix calculus is the extension of notions from calculus to matrices and vectors. We define the derivative of a scalar function $y : \mathbb{R}^n \to \mathbb{R}$ with respect to a vector $x \in \mathbb{R}^n$ as the column vector which obeys:

$$(\frac{\partial y}{\partial x})_i = \frac{\partial y}{\partial x_i}, i = 1, ...n$$

where $\frac{\partial y}{\partial x_i}$ denotes the partial derivative of y with respect to x_i. Let $A \in (R)^{n\times n}$ be an arbitrary square matrix and let $y(x) = x^T Ax$. Prove that: $\frac{\partial y}{\partial x} = (A + A^T)x$.

Solution:

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

And so, according to the product rule we get that the partial derivative with respect to x_k

would be:

$$
\begin{aligned}
\frac{\partial y}{\partial x_k} &= \frac{\partial}{\partial x_k}(\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}x_ix_j) \\
&= \frac{\partial}{\partial x_k}(x_1\sum_{i=1}^{n} a_{i1}x_1 + ... + x_k\sum_{i=1}^{n} a_{ik}x_i + ... + x_n\sum_{i=1}^{n} a_{in}x_i) \\
&= x_1a_{k1} + ... + (\sum_{i=1}^{n} a_{ik}x_i + x_ka_{kk}) + ... + x_na_{kn} \\
&= \sum_{j=1}^{n} a_{kj}x_j + \sum_{i=1}^{n} a_{ik}x_i \\
&= A_{k*}x + (A_{*k})^tx = (A_{k*} + (A_{*k})^t)x
\end{aligned}
$$

Where A_{k*} is the k-th row of A, and A_{*k} is the k-th column of A. Then:

$$
\begin{aligned}
\frac{\partial y}{\partial x} &= \begin{pmatrix} (A_{1*} + (A_{*1})^t)x \\ . \\ . \\ . \\ (A_{n*} + (A_{*n})^t)x \end{pmatrix} \\
&= \begin{pmatrix} A_{1*} + (A_{*1})^t \\ . \\ . \\ . \\ A_{n*} + (A_{*n})^t \end{pmatrix} x \\
&= \left(\begin{pmatrix} A_{1*} \\ . \\ . \\ . \\ A_{n*} \end{pmatrix} + \begin{pmatrix} (A_{*1})^t \\ . \\ . \\ . \\ (A_{*n})^t \end{pmatrix} \right) x \\
&= (A + A^t)x
\end{aligned}
$$

□

2.

Let $X_1, ..., X_n$ be i.i.d $U([0,1])$ (uniform) continous random variables. Let $Y = max(X_1, ..., X_n)$.

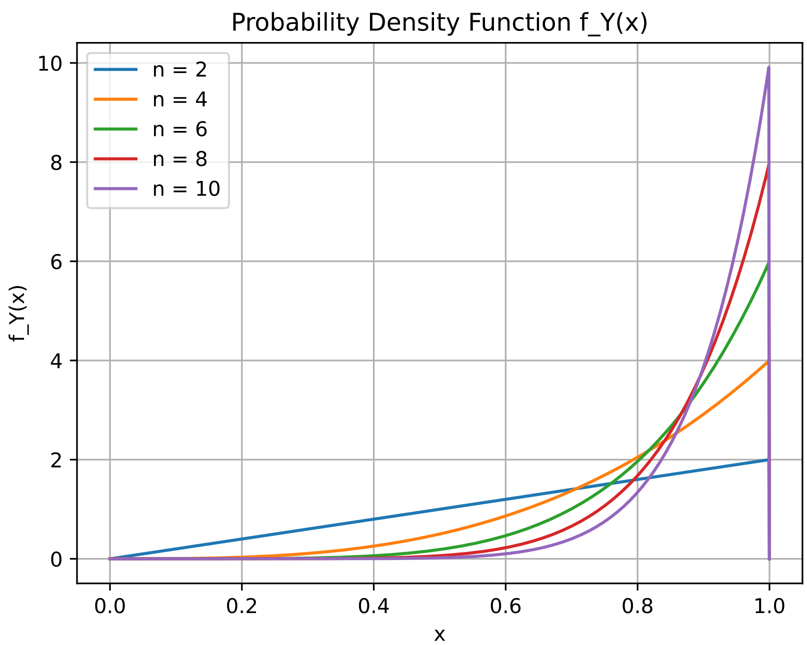
a. What is the probability density function (PDF) of Y? Write the mathematical formula and plot the PDF as well. Calculate $\mathbb{E}[Y]$ and $Var[Y]$ - how do they behave as a function of n as n grows large?

Solution: We'll calculate the CDF of Y:

$$\begin{aligned} F_Y(x) &= P[Y \leq x] = P[max(X_1, ..., X_n) \leq x] \\ &= P[X_i \leq x] = F_{X_i}(x) \cdot ... \cdot F_{X_n}(x) = (F_{X_1}(x))^n \\ &= \begin{cases} 0 & x \leq 0 \\ x^n & x \in (0,1) \\ 1 & x \geq 1 \end{cases} \end{aligned}$$

The PDF is the derivative of the CDF, so:

$$f_Y(x) = F_Y'(x) = \begin{cases} 0 & x \leq 0 \\ nx^{n-1} & x \in (0,1) \\ 0 & x \geq 1 \end{cases}$$



From here we get:

$$\mathbb{E}[Y] = \int_0^1 x \cdot nx^{n-1} dx = \int_0^1 nx^n dx = \frac{n}{n+1} = 1 - \frac{1}{n+1}$$

To calculate $Var[Y]$ we'll first calculate $\mathbb{E}[Y^2]$:

$$\mathbb{E}[Y^2] = \int_0^1 x^2 \cdot nx^{n-1} dx = \int_0^1 nx^{n+1} dx = \frac{n}{n+2}$$

And so we get:

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

As for their behavior as n grows large:

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

□

Optimal Classifiers and Decision Rules

1.

a. Let X and Y be random variables where Y can take values in $\mathcal{Y} = \{1, ..., L\}$. Let l_{0-1} be the 0-1 loss function defined in class. Show that $h = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min}\mathbb{E}[l_{0-1}(Y, f(X))]$ is given by

$$h(x) = \underset{i\in\mathcal{Y}}{\arg\max}\mathbb{P}[Y = i|X = x]$$

Solution:
Let $h(x) = \underset{i\in\{1,...,L\}}{\arg\max}P(Y = i|X = x)$. The law of total expectation states: $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$
S for any function $f : \mathcal{X} \to \{1, ..., L\}$ we'll get:

$$\begin{aligned}
\mathbb{E}[l_{0-1}(Y, f(X))] &= \mathbb{E}[\mathbb{E}[l_{0-1}(Y|f(X))|X]] \\
&= \mathbb{E}[\sum_{i=0}^{L} P(Y = i|X) \cdot l_{0-1}(f(X), i)] \\
&= \mathbb{E}[\sum_{i=0}^{L} P(Y = i|X) \cdot (1 - \mathbb{I}(f(X) = i))] \\
&= \mathbb{E}[1 - P(Y = f(X)|X)] \geq \mathbb{E}[1 - P(Y = h(X)|X)] = \mathbb{E}[l_{0-1}(Y, h(X))]
\end{aligned}$$

Where $\mathbb{I}(f(X) = i)$ is the indicator function that outputs 1 when $f(X) = i$ and 0 otherwise. The inequality follows from the fact that $h(x) = \underset{i\in\{1,...,L\}}{\arg\max}P(Y = i|X = x)$

□

b. Let X and Y be random variables where Y can take values in $\mathcal{Y} = 0, 1$. Let Δ be the following asymmetric 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 $a, b \in (0, 1]$. Compute the optimal decision rule h for the loss function Δ, i.e., the decision rule which satisfies:

$$h = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min}\mathbb{E}[\Delta(Y, f(X))]$$

Solution:
For any function $f : \mathcal{X} \to \{1, ..., L\}$ from the law of total expectation:

$$\begin{aligned}\mathbb{E}[\Delta(Y, f(X))] &= \mathbb{E}[\mathbb{E}[\Delta(Y, f(X))|X]] \\ &= \mathbb{E}[P(Y = 0|X) \cdot \Delta(0, h(x)) + P(Y = 1|X) \cdot \Delta(1, h(X))] \\ &= \mathbb{E}\left[\begin{cases} a \cdot P(Y = 0|X) & h(X) = 1 \\ b \cdot P(Y = 1|X) & h(X) = 0 \end{cases}\right]\end{aligned}$$

Then the following classifier is optimal:

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

□

2.

Let X and Y be random variables where X can take values in some set $\mathcal{X}$ and Y can take values in $\mathcal{Y} = 0, 1$. Assume 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 knows as the log-loss:

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

Find the predictor $h : \mathcal{X} \to [0, 1]$ which minimizes $\mathbb{E}[\Delta_{\log}(Y, h(X))]$.
Solution: We'll denote $\Delta_{\log} = \Delta$.

$$\begin{aligned}\mathbb{E}[\Delta(Y, h(X))] &= \sum_{x \in X, y \in Y} P(X = x, Y = y) \cdot \Delta(y, h(X)) \\ &= \sum_{x \in X} P(X = x, Y = 0) \cdot \Delta(0, h(x)) + P(X = x, Y = 1) \cdot \Delta(1, h(x)) \\ &= \sum_{x \in X} P(X = x) \left(P(Y = 0|X = x) \cdot \Delta(0, h(x)) + P(Y = 1|X = x) \cdot \Delta(1, h(x))\right) \\ &= \sum_{x \in X} P(X = x) \left(-P(Y = 0|X = x) \cdot \log(1 - h(x)) - P(Y = 1|X = x) \cdot \log(h(x))\right)\end{aligned}$$

We need to minimize: $-P(Y = 0|X = x) \cdot \log(1 - h(x)) - P(Y = 1|X = x) \cdot \log(h(x))$ We'll denote $p_0 = P(Y = 0|X = x), p_1 = P(Y = 1|X = x)$, and we'll define

$$f(x) = -p_0 \cdot \log(1 - x) - p_1 \cdot \log(x)$$

We'll differentiate f and get:

$$f'(x) = p_0 \cdot \frac{1}{1 - x} - p_1 \cdot \frac{1}{x}$$

Solving for 0:

$$\frac{p_0}{1 - x} - \frac{p_1}{x} = 0 \to \frac{p_0}{1 - x} = \frac{p_1}{x} \to p_0 \cdot x = p_1 - p_1 \cdot x \to x = \frac{p_1}{p_0 + p_1}$$

The second derivative of f is:

$$f''(x) = \frac{2p_0}{(1-x)^2} - \frac{2p_1}{x^2}$$

at $x = \frac{p_1}{p_0+p_1}$ we get

$$f''(\frac{p_1}{p_0+p_1}) = \frac{(p_0+p_1)^2}{b} + \frac{p_0}{(1-\frac{p_1}{p_0+p_1})^2} \geq 0$$

So $x = \frac{p_1}{p_0+p_1}$ is a minimum point. So we'll pick $h(x)$:

$$h(x) = \frac{P(Y=1|X=x)}{P(Y=0|X=x)+P(Y=1|X=x)} = P(Y=1|X=x)$$

□

3.

Let X and Y be random variables taking values in $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = 0, 1$ respectively, and assume that given $Y = 0$, X is distributed normally with mean μ and variance σ_0^2, i.e. $X \sim \mathcal{N}(\mu, \sigma_0^2)$, and similarly, given $Y = 1, X \sim \mathcal{N}(\mu, \sigma_0^2)$, where $\sigma_0 \neq \sigma_1$. Also assume $Pr[Y = 1] = p_1$. Find the optimal decision rule for this distribution and the zero-one loss, i.e. find $h : \mathbb{R} \to 0, 1$ which minimizes $\mathbb{E}[l_{0-1}(Y, h(X))]$ where l_{0-1} is the zero-one loss defined in class.

Solution: With Bayes' rule we get:

$$\begin{aligned}
&P(y=1|x) > P(y=0|x) \Leftrightarrow \\
&\frac{P(y=1)\cdot f_X(x|Y=1)}{f_X(x)} > \frac{P(y=0)\cdot f_X(x|Y=0)}{f_X(x)} \Leftrightarrow \\
&P(y=1)\cdot f_X(x|Y=1) > P(y=0)\cdot f_X(x|Y=0) \Leftrightarrow
\end{aligned}$$

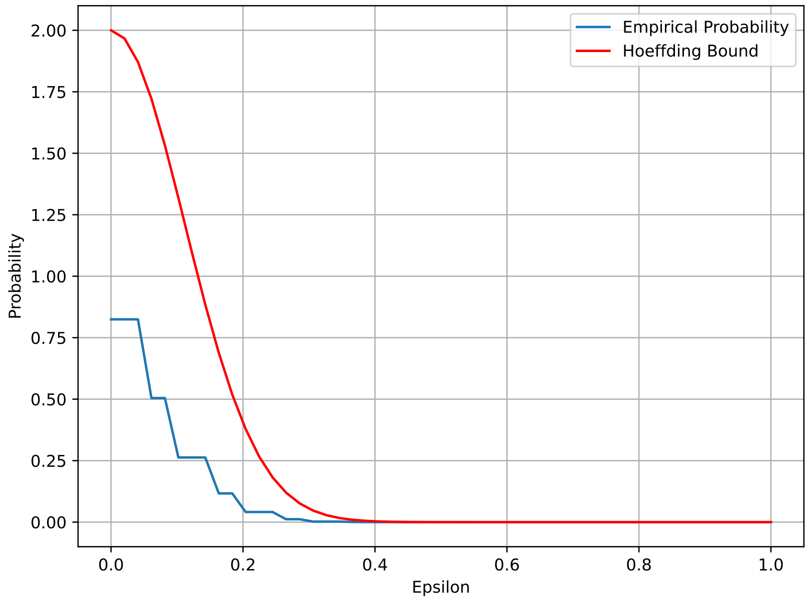
We'll denote $p = P(y = 1)$ and so $1 - p = P(y = 0)$

$$\begin{aligned}
&p \cdot f_X(x|Y=1) > (1-p)\cdot f_X(x|Y=0) \Leftrightarrow \\
&\frac{p}{\sigma_1}\cdot e^{-\frac{(x-\mu)^2}{2\sigma_1^2}} > \frac{1-p}{\sigma_0}\cdot e^{-\frac{(x-\mu)^2}{2\sigma_0^2}} \Leftrightarrow \\
&\log(\frac{p}{\sigma_1}) - \frac{(x-\mu)^2}{2\sigma_1^2} > \log(\frac{1-p}{\sigma_0}) - \frac{(x-\mu)^2}{2\sigma_0^2} \Leftrightarrow \\
&(x-\mu)^2\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right) > 2\log\left(\frac{(1-p)\sigma_1}{p\cdot\sigma_0}\right) \Leftrightarrow \\
&(x-\mu)^2(\sigma_1-\sigma_0) > \frac{2\sigma_0^2\sigma_1^2}{\sigma_1+\sigma_0}\log\frac{(1-p)\sigma_1}{p\cdot\sigma_0}
\end{aligned}$$

□

Programming Assignment

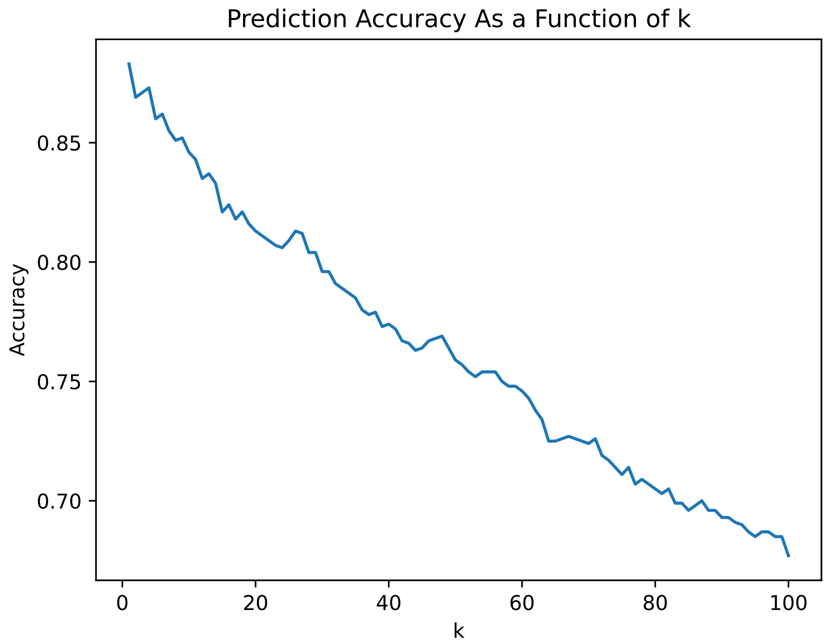
1. Visualizing the Hoeffding bound



2. Nearest Neighbor

(b) The accuracy of the prediction from the k-Nearest Neighbor algorithm with $n = 1000, k = 10$ came out to be 84.6%. Every image in the dataset is labeled from 0 to 9, then we would expect an accuracy of 10% from a completely random predictor (This is assuming a uniform distribution of labels in the dataset).

(c) Plot:



The plot shows that in general, the lower k is the better, where the best results are for $k = 1$. It seems obvious that the larger k is the less meaningful a "neighborhood" becomes, as more and more incorrect labels are included.

(d) It seems quite clear that the bigger the training set, the better the prediction gets. However, this a logarithmic plot, meaning we reach a point of diminishing returns quite quickly. And these diminishing returns also come at the cost of increase runtime.