מטלה 1 למידה חישובית

Linear Algebra

1. (**15 pts**) A symmetric matrix A over $\mathbb{R}$ is called *positive semidefinite* (PSD) if for every vector $\mathbf{v}$, $\mathbf{v}^\mathsf{T} A \mathbf{v} \geq 0$.

   (a) Show that a symmetric matrix A is PSD if and only if it can be written as $A = XX^T$, if and only if all of its eigenvalues are non-negative.

   Hint: 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.

   (b) 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 over $\mathbb{R}$ is a vector space over $\mathbb{R}$?

סעיף (a)

לשם הנוחות נגדיר:

1) $\forall v,\ v^T A v \geq 0$

2) $A = XX^T$

3) $\forall \lambda_i\ \lambda_i \geq 0$ λ_i - ע"ע של A

כעת אוכיח $2) \Rightarrow 1) \Rightarrow 3) \Rightarrow 2)$

נניח שמתקיים 2) יהי v וקטור אז:

$$v^T A v = v^T X X^T v = (X^T v)^T X^T v$$

נסמן $X^T v = y$ וקיבלנו:

$$v^T A v = \|y\|_2^2 \geq 0$$

כלומר $2) \Rightarrow 1)$.

נניח שמתקיים 1). עבור הו"ע $v_1, \dots, v_n$ והע"ע $\lambda_1, \dots, \lambda_n$ של המט' $A_{n \times n}$ מתקיים:

$$Av_i = \lambda_i v_i$$

$$v_i^T A v_i = v_i^T \lambda_i v_i = \lambda_i v_i^T v_i = \lambda_i \|v_i\|_2^2$$

אבל מכך שהנחנו ש(1) מתקיים אז $v_i^T A v_i \geq 0$ ולכן

$$\lambda_i \underbrace{\|v_i\|_2^2}_{\geq 0} \geq 0 \Rightarrow \lambda_i \geq 0$$

כלומר $(1) \Rightarrow (3)$.

נניח 3). נשתמש כרגע ונסמן $A = QDQ^T$ עבור

$$Q = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix} \qquad D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

כאשר $v_1, \dots, v_n$ ו $\lambda_1, \dots, \lambda_n$ הם ו"ע וע"ע של $A_{n\times n}$ בהתאמה.

$$A = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix}_{n\times n} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}_{n\times n} \begin{pmatrix} - & v_1 & - \\ & \vdots & \\ - & v_n & - \end{pmatrix}_{n\times n}$$

מטר' אלכסונית ולכן ניתן לכתוב אותה כמכפלה של אלכסוניות P, P^T כאשר $P = P^T$ כך:

$$D = \underbrace{\begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{pmatrix}}_{P} \underbrace{\begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{pmatrix}}_{P^T} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

הערה: בהגדרה P, P^T השתמשנו בכך שהנחנו ש(3) מתקיים ולכן $0 \leq \lambda_i$ וניתן להפעיל שורש.

ולכן בסה"כ

$$A = QPP^TQ^T = (QP)\cdot(QP)^T$$

נסמן $QP = X$ ונקבל

$$A = XX^T$$

כלומר $(3 \Rightarrow (2$.

לסיכום הוכחנו שמתקיים $(2 \Rightarrow (1 \Rightarrow (3 \Rightarrow (2$ ולכן

$$\underbrace{A \ \text{PSD}}_{(1} \Longleftrightarrow \underbrace{A=XX^T}_{(2} \Longleftrightarrow \underbrace{\forall \lambda_i,\ \lambda_i \geq 0}_{(3}$$, λ_i ע"ע של A

כנדרש.

<u>סעיף (ב)</u>

ראשית מהנתון ש A,B הן PSD נובע שהן בהכרח סימטריות לכן המטריצה $\alpha A + \beta B$ תישאר סימטרית. (הכפלה של מט' סימטרית בסקלר משאירה את המט' סימטרית וסכום של שתי מט' סימטריות $\alpha A, \beta B$ ייתן מטריצה סימטרית).

כמו כן מתקיים לכל v:

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

$$= \alpha v^TAv + \beta v^TBv$$

מכך ש A,B הן PSD אז:

$$v^TAv \geq 0,\ v^TBv \geq 0$$

מהנתון $\alpha,\beta \geq 0$ ולכן:

$$\underbrace{\alpha}_{\geq 0} \underbrace{v^TAv}_{\geq 0} + \underbrace{\beta}_{\geq 0} \underbrace{v^TBv}_{\geq 0} \geq 0$$

כלומר לכל v

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

ו $\alpha A+\beta B$ מט' סימטרית ובסה"כ היא PSD.

התוצאה שקיבלנו לא אומרת שקבוצת המטריצות
$\{A \mid A\in\mathbb{R}^{n\times n}, A \text{ is PSD}\}$ היא מרחב וקטורי מאלה או כי
קבוצה זו לא סגורה לכפל בסקלר שלילי:
לדוגמה עבור מט' A שהוא PSD ועבור $\lambda>0$ ע"ע של A
וסקלר (-1) מתקבל עבור v ו"ע של λ:

$$Av=\lambda v \Longrightarrow (-A)v=(-\lambda)v \Longrightarrow v^T(-A)v=v^T(-\lambda)v$$

$$\Longrightarrow v^T(-A)v=\underbrace{(-\lambda)}_{<0\ (*)}\ \underbrace{v^Tv}_{>0\ (**)}<0$$

ו"ע v

<u>הסברים:</u>

(*) לקחנו $\lambda>0$ ולכן $-\lambda<0$

(**) $v^Tv=\|v\|_2^2$ ולכן חיובי

כלומר קיבלנו $v^T(\alpha A)v<0$ עבור הסקלר $\alpha=-1$ ולכן
αA אינה PSD ואין סגירות לכפל בסקלר כלומר קבוצת
המטריצות הנ"ל אינה מרחב וקטורי.

Calculus and Probability

1. (**15 pts**) Let $X_1, ..., X_n$ be i.i.d $U([0,1])$ (uniform) continuous random variables. Let $Y = \max(X_1, ..., X_n)$.

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

   (b) (**No need to submit**) Verify your answer empirically using Python.

נחשב ראשית את ה CDF של Y:

$$F_Y(y) = P(Y \leq y) = P(\max(X_1, \dots, X_n) \leq y)$$

$$= P(X_1 \leq y, \dots, X_n \leq y) = \prod_{i=1}^{n} P(X_i \leq y) = \prod_{i=1}^{n} F_{X_i}(y)$$

$$= (F_{X_i}(y))^n$$

מכך ש $X_1, \dots, X_n \sim U([0,1])$ נקבל:

$$F_{X_i}(y) = \begin{cases} 0 & y < 0 \\ x & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases}$$

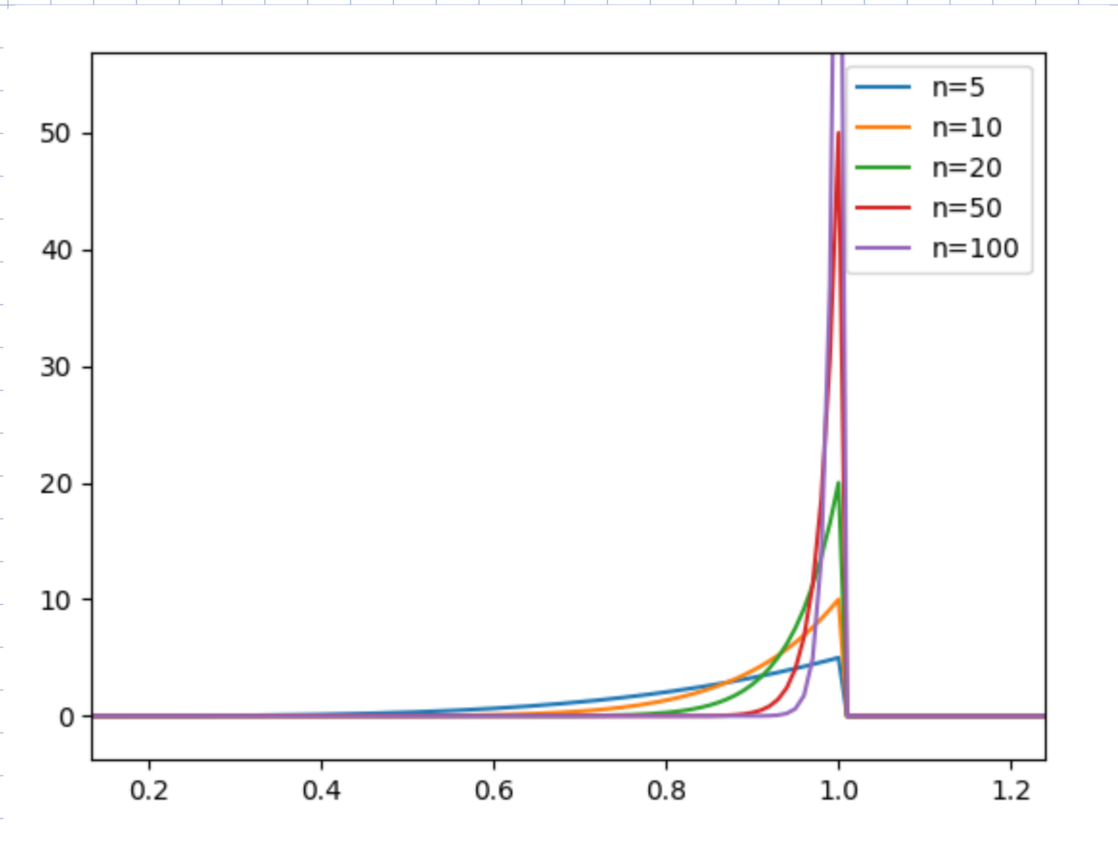
לכן:

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ x^n & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases}$$

לפי התרגול כדי לחשב PDF נצטרך לגזור את ה CDF, ונקבל:

$$f_Y(y) = \begin{cases} 0 & y < 0 \\ n y^{n-1} & 0 \leq y \leq 1 \\ 0 & y > 1 \end{cases}$$

plot:



חישוב תוחלת לפי הגדרה:

$$E[Y] = \int_0^1 y f(y)\,dy = \int_0^1 y \cdot n y^{n-1}\,dy = \int_0^1 n y^n\,dy$$

$$= n \cdot \frac{y^{n+1}}{n+1}\Big|_0^1 = n \cdot \left[\frac{1}{n+1} - \frac{0}{n+1}\right] = \frac{n}{n+1}$$

$$\lim_{n\to\infty} \frac{n}{n+1} = \lim_{n\to\infty} \frac{\frac{n}{n}}{\frac{n}{n} + \cancel{\frac{1}{n}}_{0}} = \lim_{n\to\infty} \frac{1}{1+0} = 1$$

כלומר קיבלנו $E[Y] = \frac{n}{n+1}$ ועבור n גדול(השואף לאינסוף)
התוחלת תשאף ל-1.

חישוב שונות לפי הגדרה:

$$Var(Y) = E[Y^2] - (E[Y])^2 \quad *$$

$$E[Y^2] = \int_0^1 y^2 f(y)\,dy = \int_0^1 y^2 n y^{n-1}\,dy = n\int_0^1 y^{n+1}\,dy$$

$$= n \cdot \frac{y^{n+2}}{n+2}\Big|_0^1 = n\left[\frac{1}{n+2} - \frac{0}{n+2}\right] = \frac{n}{n+2}$$

כמובן מצאנו

$$E[Y] = \frac{n}{n+1}$$

נציב ב-* ונקבל:

$$Var(Y) = \frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2 = \frac{n}{n+2} - \frac{n^2}{n^2+2n+1}$$

$$\lim_{n\to\infty}\left(\frac{n}{n+2}-\frac{n^2}{n^2+2n+1}\right)=\lim_{n\to\infty}\frac{n}{n+2}-\lim_{n\to\infty}\frac{n^2}{n^2+2n+1}$$

$$=\lim_{n\to\infty}\frac{\frac{n}{n}}{\frac{n}{n}+\underbrace{\frac{2}{n}}_{\to 0}}-\lim_{n\to\infty}\frac{\frac{n^2}{n^2}}{\frac{n^2}{n^2}+\underbrace{\frac{2n}{n^2}}_{\to 0}+\underbrace{\frac{1}{n^2}}_{\to 0}}$$

$$=\lim_{n\to\infty}\frac{1}{1}-\lim_{n\to\infty}\frac{1}{1}=1-1=0$$

כלומר קיבלנו

$$var(y)=\frac{n}{n+2}-\frac{n^2}{n^2+2n+1}$$

ועבור n גדול (שואף לאינסוף) השונות תשאף ל-0.

Optimal Classifiers and Decision Rules

1. (**15 pts**)

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

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

נרצה להביא למינימום הפסד בפונקציית ההפסד 0-1

$$\arg\min_{f:X\to y} E\left[\ell_{0-1}(Y, f(X))\right]$$

$$= \arg\min_{f:X\to y} \sum_{x,y} p(X=x, Y=y) \cdot \ell_{0-1}(Y, f(X))$$

עבור $\hat{x}$ נתון נרצה למזער את הביטויים שתלויים ב$\hat{x}$

כמו כן $y = \{1, \ldots, L\}$ ולכן:

$$\sum_{i=1}^{L} p(X=\hat{x}, y=i) \cdot \ell_{0-1}(i, f(X))$$

מהסתברות מותנה:

$$= \sum_{i=1}^{L} p(X=\hat{x}) \cdot p(y=i | X=\hat{x}) \cdot \ell_{0-1}(i, f(X))$$

$$= p(X=\hat{x}) \cdot \sum_{i=1}^{L} p(y=i | X=\hat{x}) \cdot \ell_{0-1}(i, f(X))$$

כמו כן מהגדרת פונקציית loss 0-1 מתקיים:

$$\ell_{0-1}(k,k) = 0, \quad \ell_{0-1}(k,j) = 1 \quad j \neq k$$

כלומר עבור $f(x) = i$ יתקבל $\ell_{0-1}(i, f(x)) = 0$ ועבור

$f(x) \neq i$ יתקבל $\ell_{0-1}(i, f(x)) = 1$ ולכן מתקבל:

$$p(X=\hat{x}) \cdot \sum_{i=1}^{L} p(Y=i \mid X=\hat{x}) \cdot l_{0-1}(i, f(x))$$

$$= p(X=\hat{x}) \cdot \sum_{i \neq f(x)} \overbrace{p(Y=i \mid X=\hat{x})}^{(*)}$$

נרצה לקבוע $f(x)$ שיקטין את הביטוי. נסמן $f(x)=k$ אז אם נקבע את k להיות ה-k עבורו ההסתברות $p(Y=k \mid X=\hat{x})$ היא הגדולה ביותר נקבל שהוא לא מופיע בסכום ב-(*) ולכן זה יעזור בהקטנת הביטוי. (לא נסכום את הערך הגדול ביותר). באופן פורמלי:

$$h(\hat{x}) = \arg\max_{k \in Y=\{1,\dots,L\}} p(Y=k \mid X=\hat{x})$$

(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 $a, b \in (0, 1]$ (note that this loss function generalizes the 0-1 loss defined in class). Compute the optimal decision rule h for the loss function Δ, i.e. the decision rule which satisfies:

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

$$h = \arg \min_{f:X\to Y} L(X) = \arg \min_{f:X\to Y} E\left[\Delta(Y, f(X))\right]$$

$$= \arg \min_{f:X\to Y} \sum_{x,y} P(X=x, Y=y) \cdot \Delta(y, f(x))$$

עבור $\hat{x}$ נתון נרצה למזער את הביטויים שתלויים ב-$\hat{x}$:

$$\sum_{i=0}^{1} P(X=\hat{x} \mid Y=i) \cdot \Delta(y, f(x))$$

ומתקבל:

$$= P(X=\hat{x}, Y=0) \cdot \Delta(0, f(x)) + P(X=\hat{x}, Y=1) \cdot \Delta(1, f(x))$$

ומהסתברות מותנה:

$$= P(X=\hat{x}) \cdot P(Y=0 \mid X=\hat{x}) \Delta(0, f(x))$$

$$+ P(X=\hat{x}) \cdot P(Y=1 \mid X=\hat{x}) \cdot \Delta(1, f(x))$$

$$\overset{*}{=} P(X=\hat{x}) \left[P(Y=0 \mid X=\hat{x}) \Delta(0, f(x)) + P(Y=1 \mid X=\hat{x}) \cdot \Delta(1, f(x))\right]$$

כעת נותר לבחור את $f(x)$ מתוך $0,1$. נבדוק את שני המקרים.

1) עבור $f(x)=0$ מתקיים מהגדרת פונקציית ה loss:

$$\Delta(0, f(x)) = \Delta(0,0) = 0$$

ולכן הביטוי $P(Y=0|X=\hat{x})\Delta(0,f(x))$ יתאפס.

כמו כן $\Delta(1, f(x)) = \Delta(1,0) = b$

ובסה"כ מתקבל מהצבה ב- $\circledast$:

$$\boxed{P(X=\hat{x}) \cdot P(Y=1|X=\hat{x}) \cdot b}$$

2) עבור $f(x)=1$ מתקיים מהגדרת פונקציית הloss:

$$\Delta(0, f(x)) = \Delta(0,1) = a$$

כמו כן:

$$\Delta(1, f(x)) = \Delta(1,1) = 0$$

ולכן הביטוי $P(Y=1|X=\hat{x})\cdot\Delta(1,f(x))$ יתאפס

בסה"כ מתקבל מהצבה ב $\circledast$:

$$\boxed{P(X=\hat{x}) \cdot P(Y=0|X=\hat{x}) \cdot a}$$

המטרה שלנו היא למזער את $\circledast$ ולכן נרצה לאפס את הביטוי הגדול יותר מבין הביטויים שמצאנו (מוקפים במלבן) כלומר

$$h(\hat{x}) = 1 \iff P(Y=1|X=\hat{x}) \cdot b \geq P(Y=0|X=\hat{x}) \cdot a$$

$$\iff \frac{P(Y=1|X=\hat{x})}{P(Y=0|X=\hat{x})} \geq \frac{a}{b}$$

לסיכום קיבלנו:

$$h(\hat{x}) = \begin{cases} 1 & \frac{P(Y=1|X=\hat{x})}{P(Y=0|X=\hat{x})} \geq \frac{a}{b} \\ 0 & \frac{P(Y=1|X=\hat{x})}{P(Y=0|X=\hat{x})} < \frac{a}{b} \end{cases}$$

2. (**15 pts**) 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). Assume we wish to find a predictor $h : \mathcal{X} \to [0, 1]$ (note that the hypothesis can output any number between 0 and 1) which minimizes $\mathbb{E}[\Delta_{log}(Y, h(X))]$, where Δ_{log} is the following loss function known as the *log-loss*:

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

Find the predictor $h : \mathcal{X} \to [0, 1]$ which minimizes $\mathbb{E}[\Delta_{log}(Y, h(X))]$.

Note: This loss function may seem odd at first, but it is very important and we'll discuss it further in the future.

נרצה להביא למינימום הפסד Δ_{log}:

$$\underset{h:X\to[0,1]}{\operatorname{argmin}} \mathbb{E}\left[\Delta_{log}(Y, h(X))\right]$$

$$= \underset{h:X\to[0,1]}{\operatorname{argmin}} \sum_{x,y} P(X=x, Y=y) \cdot \Delta_{log}(y, h(x))$$

עבור $\hat{x}$ נתון נרצה למזער את הביטויים שתלויים ב $\hat{x}$ כמובן

$y=\{0,1\}$ ולכן מתקבל:

$$P(X=\hat{x}, Y=0) \cdot \Delta_{log}(0, h(\hat{x})) + P(X=\hat{x}, Y=1) \cdot \Delta_{log}(1, h(\hat{x}))$$

מהסתברות מותנה:

$$(*)\ P(X=\hat{x}) \left[P(Y=0 \mid X=\hat{x}) \cdot \Delta_{log}(0, h(\hat{x})) + P(Y=1 \mid X=\hat{x}) \cdot \Delta_{log}(1, h(\hat{x}))\right]$$

מהגדרת Δ_{log} מתקבל:

$$\Delta_{log}(0, h(\hat{x})) = 0 - (1-0)\log(1 - h(\hat{x}))$$

$= -\log(1-h(\hat{x}))$

$\Delta_{\log}(1, h(\hat{x})) = -\log(h(\hat{x})) - 0 = -\log(h(\hat{x}))$

נציב ב-(*) ונקבל:

$$p(x=\hat{x})\Big[p(y=0|x=\hat{x})\cdot(-\log(1-h(\hat{x})))$$
$$+\, p(y=1|x=\hat{x})\cdot(-\log(h(\hat{x})))\Big]$$

נרצה למצוא מינימום לביטוי הזה ולכן נגזור (לפי $h(x)$) ונשווה ל-0:

$$p(x=\hat{x})\Big[p(y=0|x=\hat{x})\cdot\left(\frac{1}{1-h(\hat{x})}\right)$$
$$+\, p(y=1|x=\hat{x})\left(-\frac{1}{h(\hat{x})}\right)\Big] = 0$$

נשים לב שיתכן ש $h(\hat{x})=0$ או $h(\hat{x})=1$ ואז הביטוי לגזירה לא מוגדר. ולכן נתייחס בנפרד למקרים אלה: אם $h(\hat{x})=0$ אז h ניבא שמתקיים $p(y=0|x=\hat{x}) > p(y=1|x=\hat{x})$ ולכן הגדיר $h(\hat{x})=0$, באופן דומה עבור $h(\hat{x})=1$ אז h ניבא שמתקיים $p(y=1|x=\hat{x}) > p(y=0|x=\hat{x})$ ולכן הכריע ש $h(\hat{x})=1$. כעת נמשיך לטפל במקרה $h(\hat{x})\in(0,1)$ שגזרנו והשווינו ל-0 וכעת מתקבל:

$$p(y=0|x=\hat{x})\left(\frac{1}{1-h(\hat{x})}\right) + p(y=1|x=\hat{x})\left(-\frac{1}{h(\hat{x})}\right) = 0$$

$$\frac{p(y=0|x=\hat{x})}{1-h(\hat{x})} = \frac{p(y=1|x=\hat{x})}{h(\hat{x})}$$

$$\frac{1-h(\hat{x})}{h(\hat{x})} = \frac{P(Y=0 \mid X=\hat{x})}{P(Y=1 \mid X=\hat{x})}$$

$$\frac{1}{h(\hat{x})} - 1 = \frac{P(Y=0 \mid X=\hat{x})}{P(Y=1 \mid X=\hat{x})}$$

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

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

ומה מתכונות שמה מתקבל:

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

כלומר $h(\hat{x}) = P(Y=1 \mid X=\hat{x})$ נראה שזו אכן נק' מינימום ע"י נגזרת שנייה:

$$P(X=\hat{x})\Big[P(Y=0 \mid X=\hat{x}) \cdot \Big(\frac{1}{[1-h(\hat{x})]^2}\Big) + P(Y=1 \mid X=\hat{x}) \cdot \Big(\frac{1}{h(\hat{x})^2}\Big)\Big] > 0$$

סכום ביטויים חיוביים

ולכן אכן עבור $h(\hat{x}) = P(Y=1 \mid X=\hat{x})$ מתקבל מינימום

כלומר בסה"כ:

$$h(\hat{x}) = \begin{cases} 1 & P(Y=1 \mid X=\hat{x}) \geq P(Y=0 \mid X=\hat{x}) \\ 0 & \text{אחרת} \end{cases}$$

3. **(10 pts)**

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 $Y = 1$, $X \sim \mathcal{N}(\mu, \sigma_1^2)$, 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 $\mathbb{E}[\ell_{0-1}(Y, h(X))]$ where ℓ_{0-1} is the zero-one loss defined in class (write the decision rule only in terms of $x, \mu, \sigma_0, \sigma_1$ and p_1).

נרצה להביא למינימום את ההפסד עם פונקציית ההפסד 0-1 :

$$h = \arg\min_{h: X \to Y} E\left[\ell_{0-1}(Y, h(X))\right]$$

$$= \arg\min_{h: X \to Y} \sum_{x,y} p(X = x, Y = y) \cdot \ell_{0-1}(Y, h(X))$$

עבור $\hat{x}$ נתון נרצה למזער את הביטויים שתלויים ב-$\hat{x}$ כאשר

$y = \{0,1\}$:

$$p(X = \hat{x}, Y = 0) \cdot \ell_{0-1}(0, h(x))$$

$$+ p(X = \hat{x}, Y = 1) \cdot \ell_{0-1}(1, h(\hat{x}))$$

$$= p(X = \hat{x}) \left[p(Y = 0 \mid X = \hat{x}) \ell_{0-1}(0, h(\hat{x})) + p(Y = 1 \mid X = \hat{x}) \ell_{0-1}(1, h(\hat{x})) \right]$$

$$\ell_{0-1}(0,0) = 0 \;,\; \ell_{0-1}(1,1) = 1$$

נרצה למזער את הביטוי ולכן :

$$h(x) = \begin{cases} 1 & p(Y = 1 \mid X = \hat{x}) \geq p(Y = 0 \mid X = \hat{x}) \\ 0 & \text{אחרת} \end{cases}$$

נפסוק את. מתקיים $p(y=1|X=\hat{x}) \overset{\circledast}{\geq} p(y=0|X=\hat{x})$

כמו כן נוסחת בייס (להגדרה):

$$p(y=1|X=\hat{x}) = \frac{f_X(\hat{x}|y=1)\cdot p(y=1)}{f_X(\hat{x})}$$

$$p(y=0|X=\hat{x}) = \frac{f_X(\hat{x}|y=0)\cdot p(y=0)}{f_X(\hat{x})}$$

כלומר ⊛ שקול ל:

$$f_X(x|y=1)\cdot p(y=1) \geq f_X(x|y=0)\cdot p(y=0)$$

$$\Longleftrightarrow$$

$$\frac{f_X(x|y=1)}{f_X(x|y=0)} \geq \frac{p(y=0)}{p(y=1)}$$

$$\Longleftrightarrow$$

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

$$\Longleftrightarrow$$

$$\underbrace{\frac{\cancel{\frac{1}{\sqrt{2\pi}}}\cdot\frac{1}{\sigma_1}}{\cancel{\frac{1}{\sqrt{2\pi}}}\cdot\frac{1}{\sigma_0}}}_{\frac{\sigma_0}{\sigma_1}} \cdot e^{\left(-\frac{(x-\mu)^2}{2\sigma_1^2}+\frac{(x-\mu)^2}{2\sigma_0^2}\right)} > \frac{1-p_1}{p_1}$$

$$\Longleftrightarrow$$

$$e^{\left(-\frac{(x-\mu)^2}{2\sigma_1^2}+\frac{(x-\mu)^2}{2\sigma_0^2}\right)} > \frac{\sigma_1}{\sigma_0}\left(\frac{1-p_1}{p_1}\right)$$

הפעלת ln

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

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

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

כעת נפרק למקרים:

אם $\sigma_1^2 > \sigma_0^2$ אז $\frac{\sigma_1^2-\sigma_0^2}{\sigma_0^2\sigma_1^2} > 0$ ולכן נקבל:

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

אם $\sigma_1^2 < \sigma_0^2$ אז $\frac{\sigma_1^2-\sigma_0^2}{\sigma_0^2\sigma_1^2} < 0$ ולכן נקבל:

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

לסיכום אם $\sigma_1^2 > \sigma_0^2$ אז:

$$h(x)=\begin{cases}1 & (x-\mu)^2 > 2\ln\left(\frac{\sigma_1}{\sigma_0}\left(\frac{1-p_1}{p_1}\right)\right)\cdot\left(\frac{\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\right)\\ 0 & \text{אחרת}\end{cases}$$

אם $\sigma_0^2 > \sigma_1^2$ אז:

$$h(x)=\begin{cases}1 & (x-\mu)^2 < 2\ln\left(\frac{\sigma_1}{\sigma_0}\left(\frac{1-p_1}{p_1}\right)\right)\cdot\left(\frac{\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\right)\\ 0 & \text{אחרת}\end{cases}$$

הסברים + תשובות חלק מעשי:

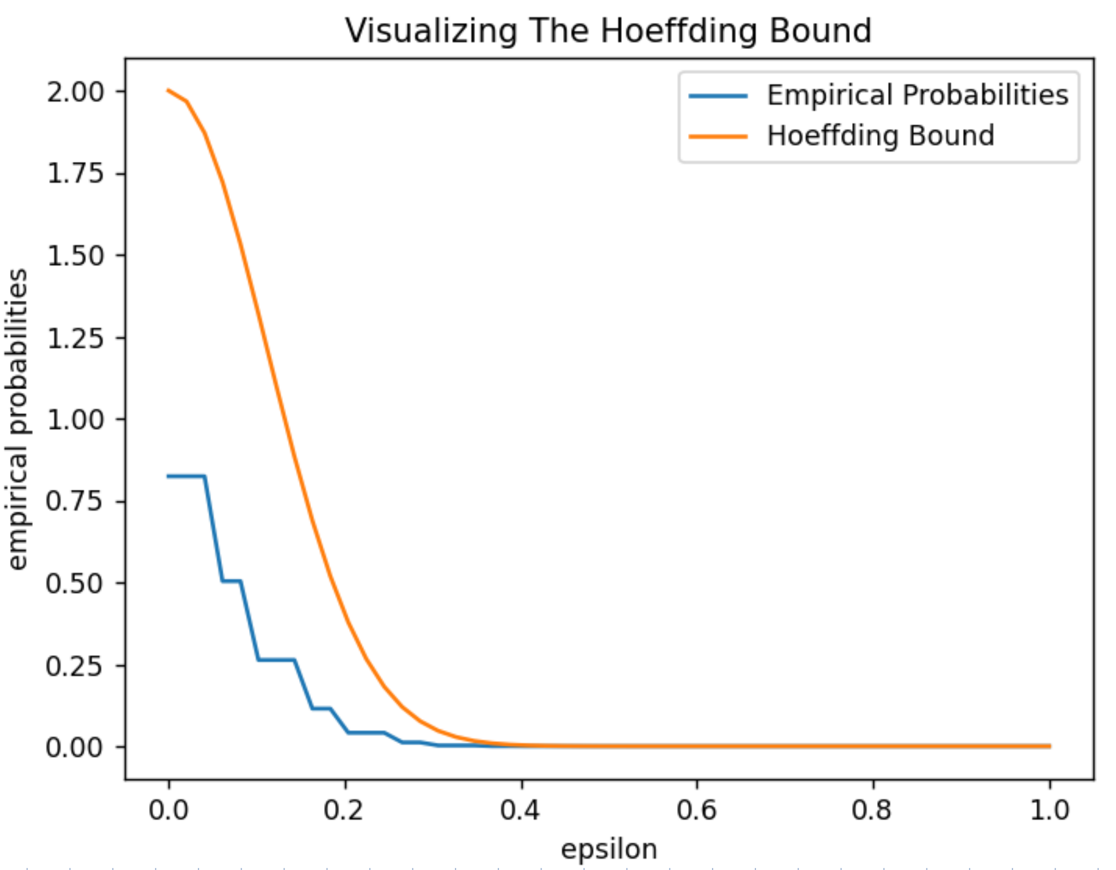
1. **Visualizing the Hoeffding bound** (10 pts).

   (a) Use **numpy** to generate an $N \times n$ matrix of samples from $Bernoulli(1/2)$. Calculate for each row i the empirical mean, $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{i,j}$, where $N = 200000$ and $n = 20$.

   (b) Take 50 values of $\epsilon \in [0,1]$ (**numpy.linspace(0,1, 50)**), and calculate the empirical probability that $|\bar{X}_i - 1/2| > \epsilon$. Plot the empirical probability as a function of ϵ.

   (c) Add to your plot the Hoeffding bound of that probability, as a function of ϵ.

   Submit your plots (no need to submit code for this question).

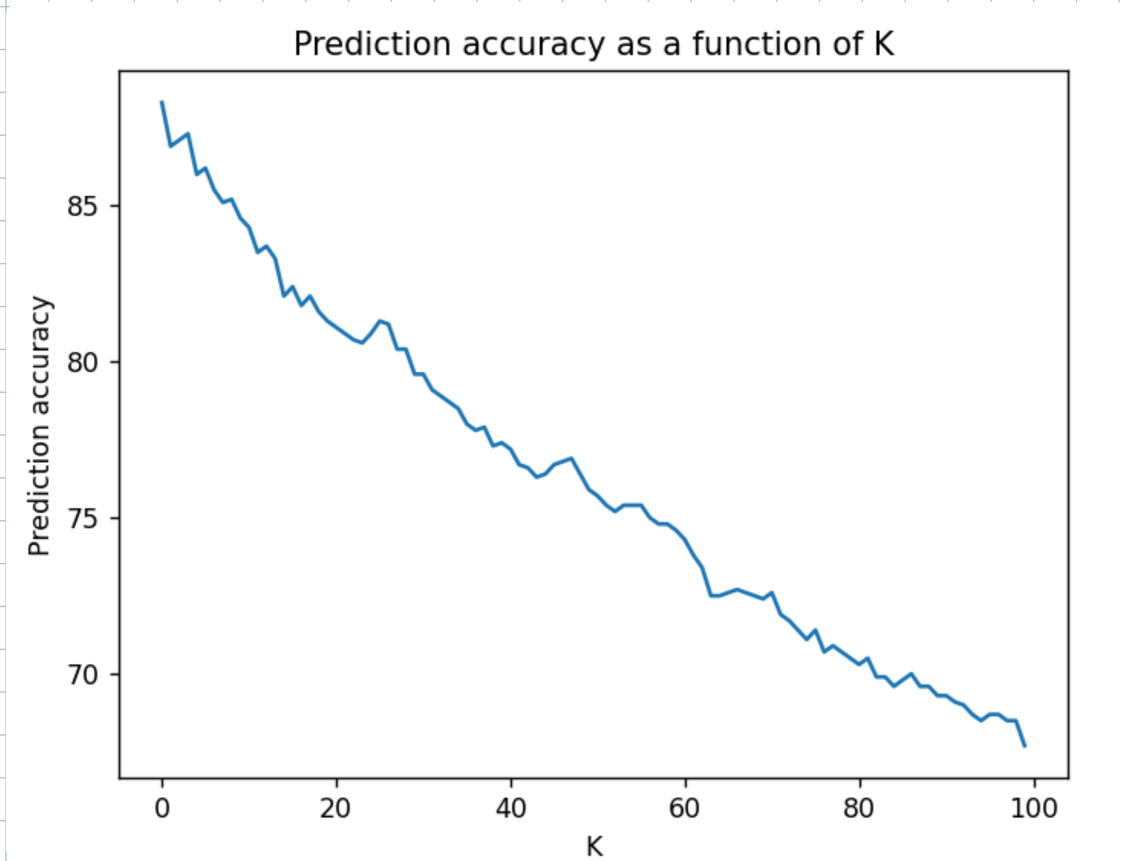


2. Nearest Neighbor (20 pts).

(b) Run the algorithm using the first $n = 1000$ training images, on each of the test images, using $k = 10$. What is the accuracy of the prediction (i.e. the percentage of correct classifications)? What would you expect from a completely random predictor?

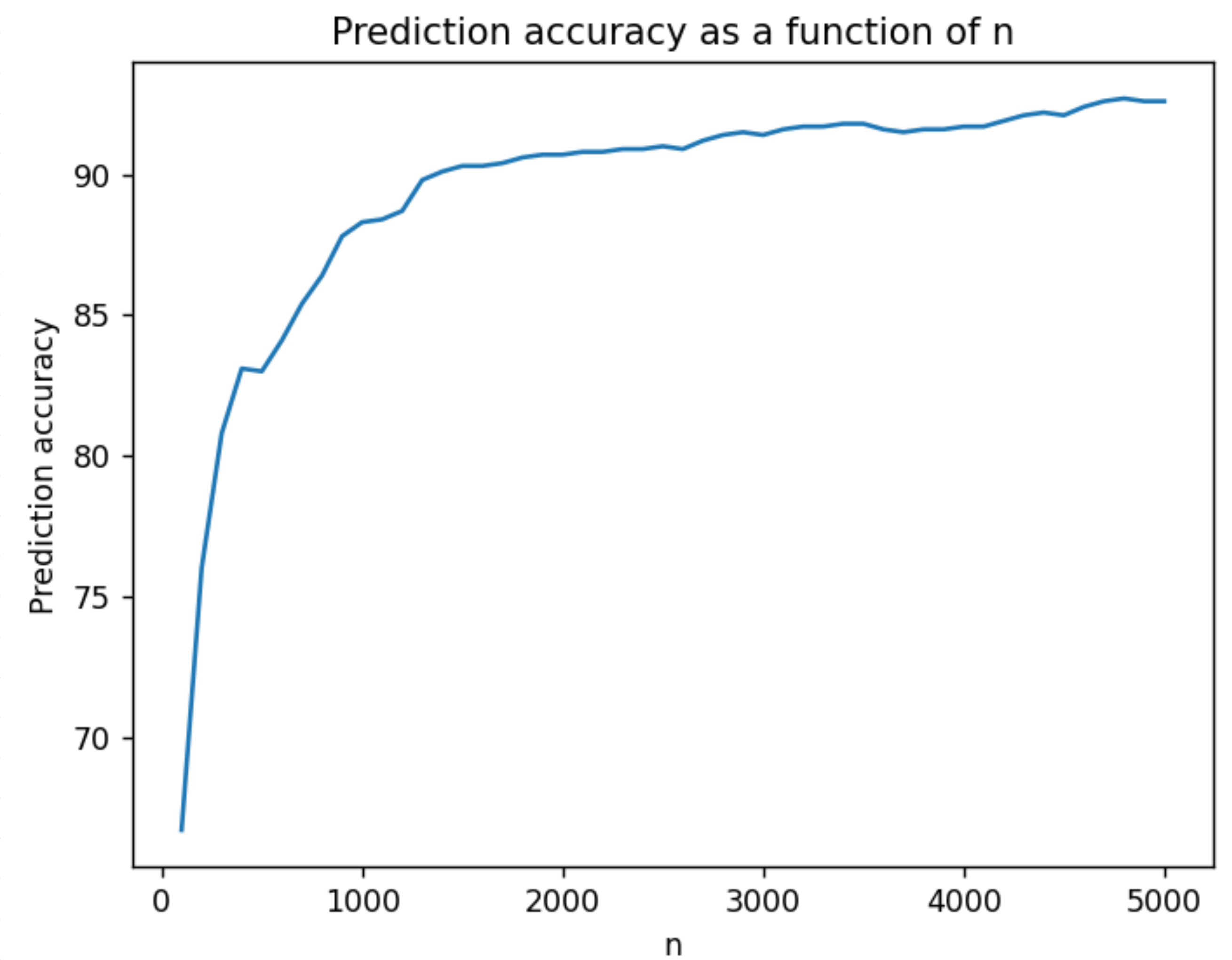
ניתן לראות בהרצת הקוד שהתקבל אחוז דיוק של 84.6%.
מפרדיקטור רנדומלי לחלוטין אצפה שאחוז הדיוק יהיה נמוך, הפרדיקטור
ינחש אקראית (באופן אחיד) מתוך ה labels 0-9 (לכן הסיכוי
שלו לצדוק יהיה $\frac{1}{10} = 10\%$

(c) Plot the prediction accuracy as a function of k, for $k = 1, \ldots, 100$ and $n = 1000$. Discuss the results. What is the best k?



ניתן לראות שככל שK גדל אחוז הדיוק הולך וקטן. נצפה לקבל אחוז דיוק גבוה ולכן הK הטוב ביותר הוא $K=1$. הסבר אינטואיטיבי לכך הוא שככל שנקח יותר שכנים יהיו יותר שכנים במרחק גדול מהתמונה שלה עושים חיזוי, ולכן הסיכוי שיהיו להם labels שונים מה label של התמונה שאותה מנסים לנבא הולך וגדל, ואיתו גם הסיכוי שנחזה label שגוי.

(d) Using $k = 1$, run the algorithm on an increasing number of training images. Plot the prediction accuracy as a function of $n = 100, 200, \ldots, 5000$. Discuss the results.



ניתן לראות שככל שn עולה אחוז הדיוק עולה, בplot הזה
הרצנו את KNN עבור K=1 ולכן ככל שנעבוד עם יותר דוגמאות
(n גדול יותר) גדל הסיכוי שנמצא שכן קרוב יותר לתמונה שעליה
עושים חיזוי, וכך גדל הסיכוי שה label של השכן יוכל
לנבא את ה label של תמונת החיזוי. כמו כן מהגרף ניתן
לראות שעבור ערכי n קטנים יותר אחוז הדיוק עולה מהר
יותר והסבר הגיוני לכך הוא כשיש כבר n גדול יותר אנו
נמצאים באחוז דיוק גבוה ולכן הוספת תמונות לא תגדיל
את אחוז הדיוק משמעותית.