IML תרגיל 3,
מגיש: אבי כוגן, ת.ז: 205417710

שאלה 1

$$h_D(x) = \underset{y\in\{\pm1\}}{argmax} Pr(y|x) \overset{Bayes\ theorem}{=} \underset{y\in\{\pm1\}}{argmax} \frac{Pr(x|y)Pr(y)}{Pr(x)} \overset{*}{=}$$
$$\underset{y\in\{\pm1\}}{argmax} Pr(x|y)Pr(y)$$

* מתקיים $Pr(x)$ לא תלוי ב-y וחיובי (x יצא לכן ההסתברות עליו אינה 0), לכן מקסימום על המונה בלבד ימקסם את הביטוי.

שאלה 2

$$\begin{aligned}
h_D(x) &= Pr(x|y)Pr(y) = \\
&= \underset{y\in\{\pm1\}}{argmax} \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} exp\{-\frac{1}{2}(x-\mu_y)^T\Sigma^{-1}(x-\mu_y)\}Pr(y) \\
&\overset{*}{=} \underset{y\in\{\pm1\}}{argmax} ln(\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} exp\{-\frac{1}{2}(x-\mu_y)^T\Sigma^{-1}(x-\mu_y)\}Pr(y)) = \\
&= \underset{y\in\{\pm1\}}{argmax} -\frac{1}{2}(x-\mu_y)^T\Sigma^{-1}(x-\mu_y) + ln(Pr(y)) = \\
&\overset{**}{=} \underset{y\in\{\pm1\}}{argmax} -\frac{1}{2}x^T\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}\mu_y + \frac{1}{2}\mu_y^T\Sigma^{-1}x - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + ln(Pr(y) = \\
&\overset{\triangle}{=} \underset{y\in\{\pm1\}}{argmax} x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + ln(Pr(y) = \underset{y\in\{\pm1\}}{argmax} \delta_y(x)
\end{aligned}$$

* ln פונק' מונוטונית עולה, לכן מקסימום על הביטוי שווה למקסימום על הln של הביטוי.
** $ln(\frac{1}{\sqrt{(2\pi)^d|\Sigma|}})$ לא תלוי ב-y ולכן ניתן
$\triangle$ $-\frac{1}{2}x^T\Sigma^{-1}x$ לא תלוי ב-y.

שאלה 3

נשתנש באומדי MLE לתוחלות ולשונות.

נחשב את $\hat{\mu}_{+1}, \hat{\mu}_{-1}, \hat{\Sigma}$: נחפש $argmaxL(\hat{\mu}_{+1}, \hat{\mu}_{-1}, \hat{\Sigma}|x_1/y_1, .., x_m/y_m)$

$$L(\hat{\mu}_{+1}, \hat{\mu}_{-1}, \hat{\Sigma}|x_1/y_1, .., x_m/y_m) = \prod_{i=1}^{m} P((x_i/y_i)/\mu_{+1}, \mu_{-1}, \Sigma) =$$

$$= \prod_{i=1}^{k} P((x_i/y_i = 1)/\mu_{+1}, \Sigma) \cdot \prod_{j=k}^{m} P((x_i/y_i = -1)/\mu_{-1}, \Sigma) =$$

$$= (\frac{1}{\sqrt{(2\pi)^d|\Sigma|}})^m exp\{\sum_{i=1}^{k} -\frac{1}{2}(x_i - \mu_{+1})^T\Sigma^{-1}(x_i - \mu_{+1})\} \cdot exp\{\sum_{j=k+1}^{m} -\frac{1}{2}(x_j - \mu_{-1})^T\Sigma^{-1}(x_j - \mu_{-1})\}$$

מתקיים: $argmax\, L(\hat{\mu}_{+1}, \hat{\mu}_{-1}, \hat{\Sigma}|x_1/y_1, .., x_m/y_m) = argmax\, ln(L(\hat{\mu}_{+1}, \hat{\mu}_{-1}, \hat{\Sigma}|x_1/y_1, .., x_m/y_m))$

,

נסמן $ln(L(\hat{\mu}_{+1}, \hat{\mu}_{-1}, \hat{\Sigma}|x_1/y_1, .., x_m/y_m)) = l(\hat{\mu}_{+1}, \hat{\mu}_{-1}, \hat{\Sigma}|x_1/y_1, .., x_m/y_m))$

$$l(\hat{\mu}_{+1}, \hat{\mu}_{-1}, \hat{\Sigma}|x_1/y_1, .., x_m/y_m) =$$

$$ln((\frac{1}{\sqrt{(2\pi)^d|\Sigma|}})^m) + \sum_{i=1}^{k} -\frac{1}{2}((x_i/y = 1) - \mu_{+1})^T\Sigma^{-1}((x_i/y = 1) - \mu_{+1}) +$$

$$\sum_{j=k+1}^{m} -\frac{1}{2}((x_j/y = -1) - \mu_{-1})^T\Sigma^{-1}((x_j/y = -1) - \mu_{-1})$$

נקבל את האומדים באמצעות גזירה והשוואה ל-0:

$$\hat{\mu}_{+1} = \frac{\partial l}{\partial \mu_{+1}} = \frac{\partial}{\partial \mu_{+1}}(-\frac{1}{2}\sum_{i=1}^{k}((x_i/y = 1)^T\Sigma^{-1}(x_i/y = 1) -$$

$$(x_i/y = 1)^T\Sigma^{-1}\mu_{+1} - \mu_{+1}\Sigma^{-1}(x_i/y = 1) + \mu_{+1}^T\Sigma^{-1}\mu_{+1}) =$$

$$= \frac{\partial}{\partial \mu_{+1}} \sum_{i=1}^{k} (-\frac{1}{2}(x_i/y = 1)^T\Sigma^{-1}(x_i/y = 1) + (x_i/y = 1)^T\Sigma^{-1}\mu_{+1} - \frac{1}{2}\mu_{+1}^T\Sigma^{-1}\mu_{+1})) =$$

$$= \sum_{i=1}^{k} ((x_i/y = 1)^T\Sigma^{-1} - \mu_{+1}^T\Sigma^{-1})$$

נשווה ל-0 למציאת האומד:

$$\sum_{i=1}^{k}((x_i/y=1)^T\Sigma^{-1}-\mu_{+1}^T\Sigma^{-1}=0\Leftrightarrow$$
$$\sum_{i=1}^{k}((x_i/y=1)^T-\mu_{+1}^T)=0\Leftrightarrow$$
$$\sum_{i=1}^{k}(x_i/y=1)=k\mu_{+1}\Leftrightarrow$$
$$\frac{1}{k}\sum_{i=1}^{k}(x_i/y=1)=\mu_{+1}$$

לכן האומד $\hat{\mu}_{+1}=\frac{1}{k}\sum_{i=1}^{k}(x_i/y=1)=\overline{(X/y=1)}\in R^d$ הוא MLE של μ_{+1}. כאשר d הוא מימד הפיצ'רים.

באופן דומה נקבל ש-$\hat{\mu}_{-1}=\frac{1}{m-k}\sum_{i=k}^{m}(x_i/y=-1)=\overline{(X/y=-1)}$

ניתן לסכם: $\hat{\mu}_{\underset{y\in\{\pm1\}}{y}}=\frac{\sum_{i=1}^{m}x_i\mathbb{I}\{y_i=y\}}{\sum_{i=1}^{m}\mathbb{I}\{y_i=y\}}$ עבור $\mathbb{I}$ פונק' אינדיקטור.

נמצא את $\hat{\Sigma}$ באמצעות גזירה לפי Σ^{-1} נציב את האומדים שממקסמים את μ_{+1},μ_{-1}:

$$\begin{aligned}\hat{\Sigma}^{-1}=\frac{\partial l}{\partial\Sigma^{-1}}(&-\frac{dm}{2}ln(2\pi)-\frac{m}{2}ln(|\Sigma|)-\frac{1}{2}\sum_{i=1}^{k}((x_i/y=1)-\overline{(X/y=1)})^T\Sigma^{-1}((x_i/y=1)-\overline{(X/y=1)})+\\&\sum_{j=k+1}^{m}-\frac{1}{2}((x_j/y=-1)-\overline{(X/y=-1)})^T\Sigma^{-1}((x_j/y=-1)-\overline{(X/y=-1)}))\\&\overset{*}{=}\frac{m}{2}\Sigma+\frac{\partial l}{\partial\Sigma^{-1}}[-\frac{1}{2}\sum_{i=1}^{k}((x_i/y=1)-\overline{(X/y=1)})^T\Sigma^{-1}((x_i/y=1)-\overline{(X/y=1)})+\\&-\frac{1}{2}\sum_{j=k}^{m}((x_j/y=-1)-\overline{(X/y=-1)})^T\Sigma^{-1}((x_j/y=-1)-\overline{(X/y=-1)}))]\\&\overset{**}{=}\frac{m}{2}\Sigma-\frac{1}{2}\sum_{i=1}^{k}((x_i/y=1)-\overline{(X/y=1)})((x_i/y=1)-\overline{(X/y=1)})^T+\\&-\frac{1}{2}\sum_{j=k}^{m}((x_j/y=-1)-\overline{(X/y=-1)})((x_j/y=-1)-\overline{(X/y=-1)})^T)\end{aligned}$$

* מתקיים $|\Sigma|=|\Sigma^{-1}|^{-1}$ לכן $ln(|\Sigma|)=-ln(|\Sigma^{-1}|)$, בנוסף $\frac{\partial l}{\partial\Sigma^{-1}}ln(|\Sigma|)=\Sigma^{-T}$ לכן נקבל $\frac{\partial l}{\partial\Sigma^{-1}}ln(|\Sigma|)=\frac{\partial l}{\partial\Sigma^{-1}}-ln(|\Sigma^{-1}|)=-\Sigma^T\overset{\Sigma\ Symmetric}{=}-\Sigma$.

** מכיוון ש-x^TAx הוא סקלר נוכל לקחת את הtr שלו ומתקיים $\frac{\partial}{\partial A}x^TAx=\frac{\partial}{\partial A}tr[x^TAx]=\frac{\partial}{\partial A}tr[Axx^T]=(xx^T)^T=xx^T$.

נשווה ל-0 למציאת האומד:

$$\frac{m}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{k}((x_i/y=1)-\overline{(X/y=1)})((x_i/y=1)-\overline{(X/y=1)})^T+$$
$$-\frac{1}{2}\sum_{j=k}^{m}((x_j/y=-1)-\overline{(X/y=-1)})((x_j/y=-1)-\overline{(X/y=-1)})^T) = 0 \Leftrightarrow$$
$$\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{k}[(x_i/y=1)-\overline{(X/y=1)})((x_i/y=1)-\overline{(X/y=1)})^T+$$
$$+\sum_{j=k}^{m}((x_j/y=-1)-\overline{(X/y=-1)})((x_j/y=-1)-\overline{(X/y=-1)})^T)]$$

קיבלנו: $\hat{\Sigma} = \frac{1}{m}\sum_{y\in\{\pm1\}}\sum_{i\in[m]\ s.t\ y_i=y}(x_i-\overline{\mu_y})(x_i-\overline{\mu_y})^T$.

ראינו בתרגול שמתקיים $\hat{Pr}_{y\in\{\pm1\}}(y) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\{y_i=y\}$

שאלה 4

הייתי מגדיר את הטעות החמורה יותר כטעות בה סיווגתי מייל שאינו ספאם כ-ספאם, מכיוון שאיבדתי מייל.
בהתאם נסווג מייל ספאם כ: 1 $Positive$, מייל לא ספאם כ: 1- ,$Negative$.
טעות מסוג ראשון, חמורה יותר - FP - סיווגתי מייל שאינו ספאם כ-ספאם. (כלומר איבדתי מייל אמיתי).
טעות מסוג שני, פחות חמורה - FN -סיווגתי מייל ספאם כלא ספאם. (קיבלתי בטעות ספאם).

שאלה 5

נגדיר $\begin{pmatrix} w \\ b \end{pmatrix} = v$ בהתאם לנוסחה ונקבל:

$$\underset{\substack{(w,b) \\ s.t\ \forall i,\ y_i(<w,x_i>+b)\geq 1}}{argmin}||w||^2 \overset{*}{=}$$
$$\underset{(w,b)}{argmin}\begin{pmatrix} w \\ b \end{pmatrix}^T I \begin{pmatrix} w \\ b \end{pmatrix} =$$
$$s.t \begin{pmatrix} (y_1x_1) & y_1 \\ . & . \\ . & . \\ (y_mx_m) & y_m \end{pmatrix}\begin{pmatrix} w \\ b \end{pmatrix} \geq \begin{pmatrix} 1 \\ . \\ . \\ 1 \end{pmatrix}$$
$$\underset{(w,b)}{argmin}\frac{1}{2}\begin{pmatrix} w \\ b \end{pmatrix}^T 2I \begin{pmatrix} w \\ b \end{pmatrix}$$
$$s.t -\begin{pmatrix} (y_1x_1) & y_1 \\ . & . \\ . & . \\ (y_mx_m) & y_m \end{pmatrix}\begin{pmatrix} w \\ b \end{pmatrix} \leq \begin{pmatrix} 1 \\ . \\ . \\ 1 \end{pmatrix}$$

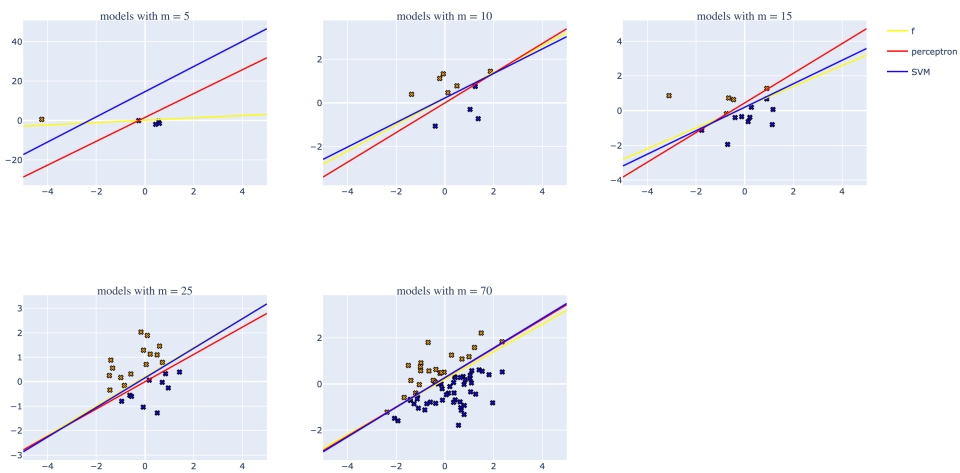
לכן קיבלנו $d = \begin{pmatrix} 1 \\ \cdot \\ \cdot \\ 1 \end{pmatrix}$, $a = 0$, $Q = 2I$, $A = \begin{pmatrix} (y_1 x_1) & y_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ (y_m x_m) & y_m \end{pmatrix}$.

שאלה 6

ניתן להמיר את הבעיה לצורה האלטרנטיבית מכיוון שמציאת מינ' על פני w, ξ_i שקול, נראה זאת בכך שעבור w מסויים המני' של ξ_i תחת האילוצים שהוא אי שלילי וצריך להתקיים $y_i < w, x_i > \geq 1 - \xi_i$ נקבל שבמידה ו-$y_i < w, x_i > \geq 1$ ההשמה האופט' עבור ξ_i היא $\xi_i = 0$. אבל עבור המקרה $y_i < w, x_i > < 1$ ההשמה הטובה ביותר היא $\xi_i = 1 - y_i < w, x_i >$ על מנת לשמר את האילוץ בצורה ההדוקה ביותר. לכן נקבל שהצורה האלטרנטיבית היא שקולה לצורה המקורית מכיוון שפונק' l^{hing} מביאה למינ' את ξ_i תחת האילוצים שלו.

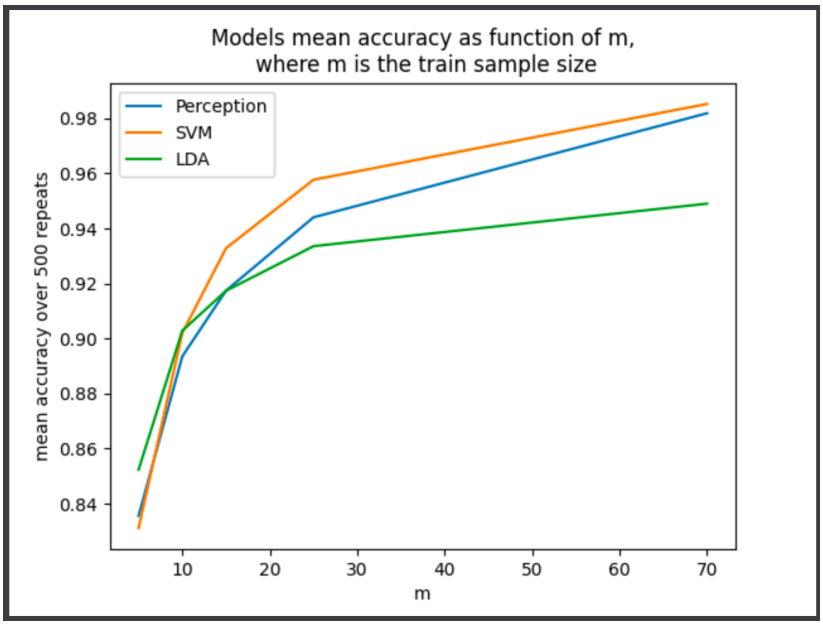
שאלה 9

תרשים פיזור נקודות מוגרלות עבור מדגם אימון בגודל m, יחד איתו על-המישור המתאים לכל מודל.



שאלה 10

התרשים:

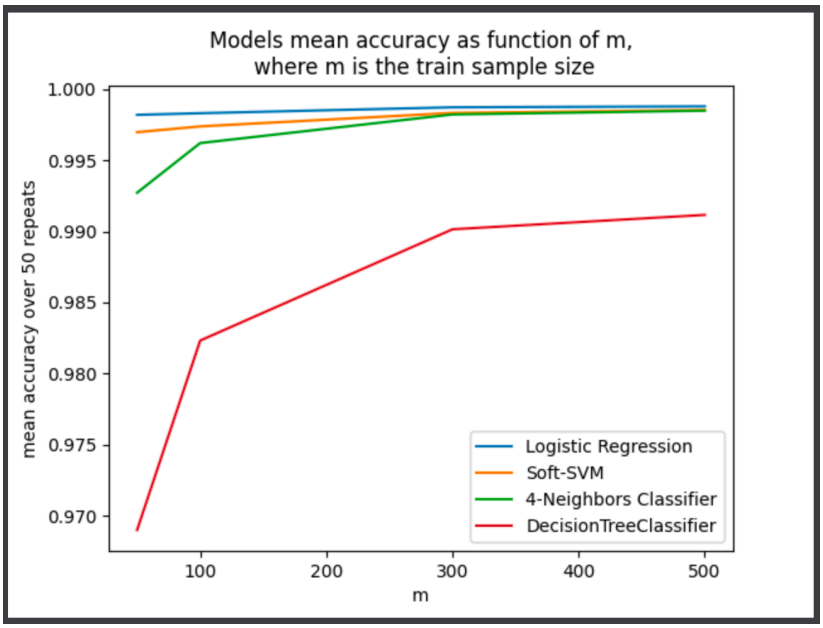


שאלה 11

ניתן לראות ש-SVM הצליח יותר, זאת מכיוון שהוא מתאים את עצמו בצורה טובה יותר כאשר ממקסם את ה-$margin$ בשונה מה-$Perceptron$ שמתאים את הקו הראשון שמצליח לחלק את הדאטה. ה-LDA הצליח פחות משניהם מכיוון שמניח שהדאטה שמייצר את 2 הלייבל מגיע מהתפלגות נורמלית שונה כל אחד, אך במקרה זה הוא מגיע מהתפלגות יחידה.

שאלה 14

התרשים:



כאשר ארבעת המודלים מומשו בעזרת ספריית $sklearn$,
עבור $Soft-SVM$ בחרתי בפרמטר רגולציה 1,
עבור $kNeighborsClassifier$ בחרתי ב-4 שכנים,
עבור $DecisionTreeClassifier$ בחרתי בעומק 4 מקסימלי.
זמני הריצה הם:
עבור $LogisticRegression$: 5.23 שניות
עבור SVC: 9.51 שניות
עבור $4-NeighborsClassifier$: 17.60 שניות
עבור $DecisionTreeClassifier$: 1.87 שניות

ניתן לראות שלעץ עם עומק 4 לקח הכי מעט זמן מכיוון שהאימון שלו מהיר וגם הסיווג בעומק שכזה.
לעומת זאת מודל $4-NeighborsClassifier$ לקח הכי הרבה זמן וזאת מכיוון שתהליך סיווג כל תצפית אינו פשוט בהשוואה לאחרים.
ל-SVC לקח יותר זמן בהשוואה ל $LogisticRegression$, יתכן שבעקבות תהליך האימון שדורש מציאת על-מישור.