Theoretical Assignment 3

Intro to Machine Learning, TAU

December 17, 2024

Question 1

The Hyperbolic Tangent kernel is defined as $k(x, y) = tanh(x^T y + \gamma)$. Where $\gamma \in \mathbb{R}$ is some constant scalar.
Show that this is not a valid kernel according to Mercer's theorem.
Hint: Develop the conditions for a set of 2 samples: $S = \{x_1, x_2\}$.

Question 2

The hinge loss $\max(0, 1 - y(w^T x + b))$ is commonly used in SVMs. However, consider an alternative quadratic loss:

$$L_i(w, b) = \left[\max(0, 1 - y_i(w^T x_i + b))\right]^2.$$

(a) In soft-SVM, what do the slack variables represent? What is the geometrical meaning of the different value ranges that the slack variables can have? How are the slack variables related to the hinge loss?

(b) Write the primal optimization problem for a soft-margin SVM using this new quadratic loss. Hint: Consider about your answers to section a to understand how it would be different from the primal problem with the hinge loss.

(c) Write the Lagrangian, KKT conditions, and the dual problem for a soft-margin SVM using this new quadratic loss.

(d) Derive the gradient of the loss function $L_i(w, b)$ with respect to w and b.

(e) Discuss the implications of using this loss function instead of the hinge loss. How might this impact the sparsity of support vectors and the generalization performance?

Question 3

(a) What is the sample complexity of hard margin SVM in the linear case (without a kernel) and why? Explain your answer, don't just quote from class.

(b) Briefly explain how will using an RBF kernel affect the sample complexity. For an SVM using the RBF kernel, discuss how the sample complexity is influenced by the kernel hyperparameter σ, which controls the spread of the RBF kernel.

Question 4

Given a dataset with two classes, C_1 and C_2, and the following class distributions:

Feature Value	Class C_1 Count	Class C_2 Count
A	8	2
B	3	7

(a) Compute the entropy $H(S)$ of the entire dataset and for the sunsetes S_A S_B (the subsets corresponding to feature values A and B).
(b) Calculate the information gain $G(S, A)$ when splitting the dataset on the feature value.
(c) Interpret your results: What does the information gain tell you about the usefulness of splitting on this feature?
(d) Compute the Gini impurity for the entire dataset S, and for the subsets S_A and S_B **before** splitting.
(e) Calculate the weighted average Gini impurity **after** splitting the dataset by the feature values A and B.
(f) Compare the results from entropy-based information gain and Gini impurity. Are the rankings of the splits the same? Explain why or why not.

Question 5

Consider a binary decision tree used for binary classification, such as the ones we've seen in class, where each internal node splits the data based on a single feature threshold, and each leaf represents a class label. Suppose the dataset has n features, and each feature can take on real values.
(a) Prove that a binary decision tree with depth d can divide the feature space into at most 2^d regions.
(b) Explain how this result relates to the expressiveness of decision trees and their capacity to overfit.
(c) Using this result, determine the minimum depth d required for a decision tree to perfectly classify a dataset of m unique points (no two points share the same feature values).
(d) Based on the previous sections, what is the VC dimension of a tree with depth d?

Question 1

The Hyperbolic Tangent kernel is defined as $k(x, y) = tanh(x^T y + \gamma)$. Where $\gamma \in \mathbb{R}$ is some constant scalar.
Show that this is not a valid kernel according to Mercer's theorem.
Hint: Develop the conditions for a set of 2 samples: $S = \{x_1, x_2\}$.

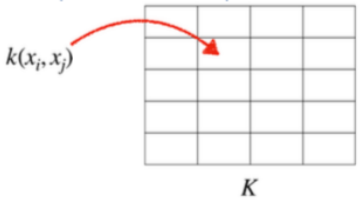
$m = 2$

$S = \{x_1, x_2\}$ $\quad x_{1,2} \in \mathbb{R}^d$

$k_{ij} = k(x_i, x_j)$

$K = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) \\ k(x_2,x_1) & k(x_2,x_2) \end{pmatrix}$

- **Mercer's theorem**



A function $k : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is a kernel iff

- For any $m \in \mathbb{N}$ and any $x_1, \ldots, x_m \in \mathbb{R}^d$.
- Define the matrix $K \in \mathbb{R}^{m \times m}$ where $K_{i,j} = k(x_i, x_j)$ for all $i, j = 1, \ldots, m$.
- Then K is symmetric and positive semi-definite.

K is positive semidefinite

$\forall w \neq 0 \quad w^T K w \geq 0$

$w = \begin{pmatrix} a \\ b \end{pmatrix}$, $w \neq 0$, $w \in \mathbb{R}^2$ [illegible]

$w^T K w < 0$

positive semidefinite if K [illegible]

$$\begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a k_{11} + b k_{21} & a k_{12} + b k_{22} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} =$$

$$a^2 \tanh\left(\|x_1\|^2 + \beta\right) + 2ab \tanh\left(x_1^T x_2 + \beta\right) + b^2 \tanh\left(\|x_2\|^2 + \beta\right)$$

$b = -1$, $a = 1$ נקבל

$$a^2 \tanh\left(\|x_1\|^2 + \beta\right) - 2 \tanh\left(x_1^T x_2 + \beta\right) + \tanh\left(\|x_2\|^2 + \beta\right)$$

נגדיל את:

$$x_1 = \begin{pmatrix} x_1 \\ x_1 \end{pmatrix} , \quad x_1 \to \infty$$

$$x_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \rightsquigarrow \|x_2\| = 1$$

$$\tanh(\infty) - 2\tanh(\infty) + \tanh(1+\beta) = -1 + \tanh(1+\beta) < 0$$

$\beta \in \mathbb{R}$ $\Rightarrow$ $-\infty < 1 + \beta < \infty$

$\Downarrow$

$-1 < \tanh(1+\beta) < 1$

Question 2

The hinge loss $\max(0, 1 - y(w^T x + b))$ is commonly used in SVMs. However, consider an alternative quadratic loss:

$$L_i(w, b) = \left[\max(0, 1 - y_i(w^T x_i + b))\right]^2.$$

(a) In soft-SVM, what do the slack variables represent? What is the geometrical meaning of the different value ranges that the slack variables can have? How are the slack variables related to the hinge loss?

(b) Write the primal optimization problem for a soft-margin SVM using this new quadratic loss. Hint: Consider about your answers to section a to understand how it would be different from the primal problem with the hinge loss.

(c) Write the Lagrangian, KKT conditions, and the dual problem for a soft-margin SVM using this new quadratic loss.

(d) Derive the gradient of the loss function $L_i(w, b)$ with respect to w and b.

(e) Discuss the implications of using this loss function instead of the hinge loss. How might this impact the sparsity of support vectors and the generalization performance?

$\varepsilon_i > 1$ הנקודה לא מסווגת נכון

* נשים' ה- slack פשוט hinge lost -

מ- slackness complementary (KKT) נקבל

כולם לוקחים הדוקים $y_i(w^T x_i + b) = 1 - \varepsilon_i$

$$\Downarrow$$

$$\varepsilon_i = \max\{0, 1 - y_i(w^T x_i + b)\}$$

$$\Updownarrow$$

פשוט ה- hinge lost

b)

(a- מ) לכן הבעיה Soft SVM של hinge lost

$$\min_{w,b,\varepsilon} \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i$$

$$\Downarrow$$

$$\varepsilon_i = hinge(w, b, (x_i, y_i))$$

עבודה פונקציית ההפסד בשאלה

$$L_i(w,b) = [\ell_{hinge}(w,b(x_i,y_i))]^2 = \varepsilon_i^2$$

• אין צורך להביא את ההפסד כך — למינימום את לבדו

למינימום את ε_i^2

• ε_i עדיין מייצג את — הגבול של המרווח

לכן בעיית האופטימיזציה נשארת אותו דבר

בעיית האופטימיזציה החדשה תחת שינוי ההפסד

הבעיה הסופית

$$\min_{w,b,\varepsilon} \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2$$

$$y_i(w^T x_i + b) \ge 1-\varepsilon_i \qquad \varepsilon_i \ge 0 \quad \forall i$$

- בעיית האופטימיזציה הזו, בהשוואה לבעיית האופטימיזציה עם פונקציית ה- hinge "מענישה יותר" כאשר ε_i גדול

ג) הצגת הבעיה כבעיית אופטימיזציה ללא אילוצים

$$L(w,b,\xi,\alpha,\mu) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\left[y_i(w^T x_i + b) - 1 + \xi_i\right] - \sum_{i=1}^{n}\mu_i\xi_i$$

$$L(w,b,\xi,\alpha,\mu) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\left(\alpha_i y_i w^T x_i + \alpha_i y_i b - \alpha_i + \alpha_i\xi_i\right) - \sum_{i=1}^{n}\mu_i\xi_i$$

$$\frac{\partial L}{\partial w} = \lambda w - \sum_{i=1}^{n}\alpha_i x_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = \frac{2}{n}\xi_i - \alpha_i - \mu_i = 0$$

$$\frac{\partial L}{\partial w} = -\sum_{i=1}^{n}\alpha_i y_i = 0$$

$$\begin{cases} w = \frac{1}{\lambda}\sum_{i=1}^{n}\alpha_i x_i y_i \\ \sum_{i=1}^{n}\alpha_i y_i = 0 \\ \xi_i = \frac{n}{2}(\alpha_i + \mu_i) \end{cases}$$

$y_i(w^T x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$ $\forall i$ — primal feasibility

$\alpha_i \geq 0$, $\mu_i \geq 0$ $\forall i$ — dual feasibility

$\alpha_i(1 - \xi_i - y_i(w^T x_i + b)) = 0$ $\mu_i \xi_i = 0$ $\forall i$ — complementary slackness

נפתח את בדיקה המקרים.

מתנאי complementary slackness ($\mu_i \xi_i = 0$)

$\mu_i = 0$ או $\xi_i = 0$ שניהם מתקיים,

ולכן אם $\xi_i > 0$ ← $\mu_i = 0$

כלומר מכאן המשוואות שקיבלנו

משמעותי הפתרונות, עבור $\xi_i = \frac{n}{\lambda} \alpha_i$

ונציב בדיקה הפתרונות בכל ה—

הביטוי לחזור את הדואלי:

$$\|w\|^2 = \left\|\frac{1}{\lambda} \sum_{i=1}^{n} \alpha_i x_i y_i\right\|^2 = \frac{1}{\lambda^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$ (חלק 1)

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{n}{2}\alpha_i\right)^2=\frac{n}{4}\sum_{i=1}^{n}\alpha_i^2$$

סה״כ נקבל הדואלי -

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i^T x_j)-\frac{n}{4}\sum_{i=1}^{n}\alpha_i^2$$

$$\text{s.t.}\ \sum_{i=1}^{n}\alpha_i y_i=0 \qquad \alpha_i\geq 0$$

ד) נחשב את הגרדיאנט לפי שני המקרים הבאים.

① $1-y_i(w^Tx_i+b)<0$: $L_i=0$

$$\frac{\partial L_i}{\partial w}=0 \qquad \frac{\partial L_i}{\partial b}=0$$

② $1-y_i(w^Tx_i+b)\geq 0$: $L_i=(1-y_i(w^Tx_i+b))^2$

$$L_i=1-2y_i(w^Tx_i+b)+y_i^2(w^Tx_i+b)^2=$$

$$1-2y_iw^Tx_i-2y_ib+y_i^2((w^Tx_i)^2+2w^Tx_ib+b^2)=$$

$$1-2y_iw^Tx_i-2y_ib+y_i^2(w^Tx_i)^2+y_i^2(2w^Tx_ib)+y_i^2b^2$$

$$\frac{\partial L_i}{\partial w}=-2y_ix_i+2y_i^2w^Tx_ix_i+2y_i^2x_ib=2y_ix_i(-1+y_iw^Tx_i+y_ib)$$

$$=-2y_ix_i(1-(y_i)(w^Tx_i+b))$$

$$\frac{\partial L_i}{\partial b} = -2y_i + 2y_i^2 w^T x_i + 2y_i^2 b = -2y_i(1-(y_i)(w^T x_i + b))$$

e) ראינו שבניגוד ל- hinge loss הפונקציה square מענישה "יותר" לדוגמאות עם slack גדול יותר כלומר רגישה יותר לנקודות חריגות יותר, כמה שהוא יותר רחוק (קיצוני) מאזור המידע עליו מענישים את המודל, ישפיעו על קו ההחלטה (לא רק ה-SV)

לכן ההפסד הריבועי יותר רגיש לקיצוניות גם שהוא בצד הנכון של קו ההחלטה שהיא, תוצר יותר משפיע להשוואה של קיצוני- הגדול. כך יגרום לשינוי של קו ההחלטה לקיצוני.

בניגוד לא קיים הפסד הריבועי יותר. וזה בניגוד לפונ' ה- hinge שבה הענישה לדוגמה המרוחק מקו ההחלטה הינו לינארי. זה עלול לגרום ל- variance יותר גבוה ולסיכויים יותר גבוה ל- overfitting.

לכן פונק' ההפסד הריבועי — שמה LL2
על הפרט "רעש" יש ω כלל ההחלטה,
מה שהופך את — המודל לרגיש יותר עקבי ?
לא כמו פונק' ה- שגיאה שהיא לינארית —
במרחק ω הקרוב לכלל ההחלטה
כלומר שימוש בפונק' הטעות יהיה סימן יותר
טובה ל- overfitting כי המודל יהיה רגיש
יותר לכל הוקטור G ω את — מהנקודות.

Question 3

(a) What is the sample complexity of hard margin SVM in the linear case (without a kernel) and why? Explain your answer, don't just quote from class.

(b) Briefly explain how will using an RBF kernel affect the sample complexity. For an SVM using the RBF kernel, discuss how the sample complexity is influenced by the kernel hyperparameter σ, which controls the spread of the RBF kernel.

(a)

הסיווגים — (קבוע — זה ה- VC-dim של ה- hypothesis class
שהוא שווה למספר הנקודות d שנוכל לנפץ

hard SVM מוצא את — ה- hyperplane
שממקסם המרחק בין ה-class-ים שמופרדים לינארית —
מה- hyperplane

$\forall i \quad w^T x_i + b = 0$

bias שמזיז את — ה- hyperplane
w - המשקל של הוקטור b ו
ביחד — מגדירים אותו
.לו

ה- Sample complexity הוא מספר הפרמטרים של ה- seperator
שלנו בדיוק ה..., d
d+1 -

$$err(h_{\hat{w},\hat{b}}) = O\left(\frac{d+1+\log\frac{1}{\delta}}{n}\right)$$

לכן

$$n = O\left(\frac{d+1+\log\frac{1}{\delta}}{\varepsilon}\right)$$

כי שגיאת ההכללה היא ε שקטנה דרסטית

d לא גדול ב־ המימד $\mathcal{D}$ הקלט במימד d, אלא

כגודל של מודל ה־ VC-dim,

(b) RBF kernel ממפה את הקלט ← ϕ למימד גבוה מאוד

במימד גבוה כלומר, ψ, ה־sample complexity גבוה מאוד

כי נצטרך הרבה data כדי שלא יהיה overfit.

יש גם kernel hyperparameter, σ,

$$k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

כאשר $x,y \in \mathbb{R}^d$

ניתן לראות שכש-γ (גדול ה-σ קטן) kernel
הרבה יותר פשוט שיהיה מפריד יותר רחוק, והמודל
שלא יהיה פשוט יותר וצריך פחות דגימות
ופחות samples כמות שיכול להוביל overfitting
(דוגמה הבאה)

אם γ קטן — σ אז ה RBF Kernel יהיה
מאוד ממוקד כי המודל יצטרך להבחין יותר
הדגימות שקרובות השוואה קרובות לא לא
ופחות יתייחס לדגימות הרחוקות,
כלומר, המודל יהיה יותר מסובך ויותר להתאים
overfitting, אבל צריך יותר דגימות.

Question 4

Given a dataset with two classes, C_1 and C_2, and the following class distributions:

Feature Value	Class C_1 Count	Class C_2 Count
A	8	2
B	3	7

(a) Compute the entropy $H(S)$ of the entire dataset and for the sunsetes S_A S_B (the subsets corresponding to feature values A and B).
(b) Calculate the information gain $G(S, A)$ when splitting the dataset on the feature value.
(c) Interpret your results: What does the information gain tell you about the usefulness of splitting on this feature?
(d) Compute the Gini impurity for the entire dataset S, and for the subsets S_A and S_B **before** splitting.
(e) Calculate the weighted average Gini impurity **after** splitting the dataset by the feature values A and B.
(f) Compare the results from entropy-based information gain and Gini impurity. Are the rankings of the splits the same? Explain why or why not.

a) $H(S) = -(s\log_2(s) + (1-s)\log_2(1-s))$ $|S| = 20$

$= -\left(\frac{11}{20}\log_2\left(\frac{11}{20}\right) + \left(\frac{9}{20}\right)\log_2\left(\frac{9}{20}\right)\right) = 0.992$

$H(S_A) = -\left(\frac{8}{10}\log_2\left(\frac{8}{10}\right) + \left(\frac{2}{10}\right)\log\left(\frac{2}{10}\right)\right) = 0.722$ $|S_A| = 10$ $:S_A$

$H(S_B) = -\left(\frac{3}{10}\log_2\left(\frac{3}{10}\right) + \left(\frac{7}{10}\right)\log\left(\frac{7}{10}\right)\right) = 0.881$ $|S_B| = 10$ $:S_B$

b) information gain

A עמ' (30)

$$G(S,A) = H(S) - H(A) = ?$$

$$H(A) = P_r(A) \cdot H(S_A) + P_r(B) \cdot H(S_B) =$$

$$\frac{10}{20} \cdot 0.722 + \frac{10}{20} \cdot 0.881 = 0.802$$

$$\Downarrow$$

$$G(S,A) = H(S) - H(A) = 0.19$$

c) [illegible]

d) Gini index $= Var(q) = q(1-q)$

$$H(S) = 2S(1-S) = 2 \cdot \frac{11}{20}\left(\frac{9}{20}\right) = 0.495$$

$$H(S_A) = 2 \cdot \frac{8}{10} \cdot \frac{2}{10} = 0.32$$

$$H(S_B) = 2 \cdot \frac{3}{10} \cdot \frac{7}{10} = 0.42$$

e) $H(\text{after split}) = Pr(A) H(S_A) + Pr(B) H(S_B) =$

$\frac{1}{3} \cdot 0.32 + \frac{2}{3} \cdot 0.42 = 0.37$

f)

Question 5

Consider a binary decision tree used for binary classification, such as the ones we've seen in class, where each internal node splits the data based on a single feature threshold, and each leaf represents a class label. Suppose the dataset has n features, and each feature can take on real values.

(a) Prove that a binary decision tree with depth d can divide the feature space into at most 2^d regions.

(b) Explain how this result relates to the expressiveness of decision trees and their capacity to overfit.

(c) Using this result, determine the minimum depth d required for a decision tree to perfectly classify a dataset of m unique points (no two points share the same feature values).

(d) Based on the previous sections, what is the VC dimension of a tree with depth d?

a)

Regions $\rightsquigarrow$ 2^d

R-Regions

$d=0$

$R(0)=2^0=1$

Split

צעד - נניח שלכל עץ בעומק d-1

$$R(d-1)=2^{d-1}$$

(כלומר לכל היותר 2^{d-1} עלים)

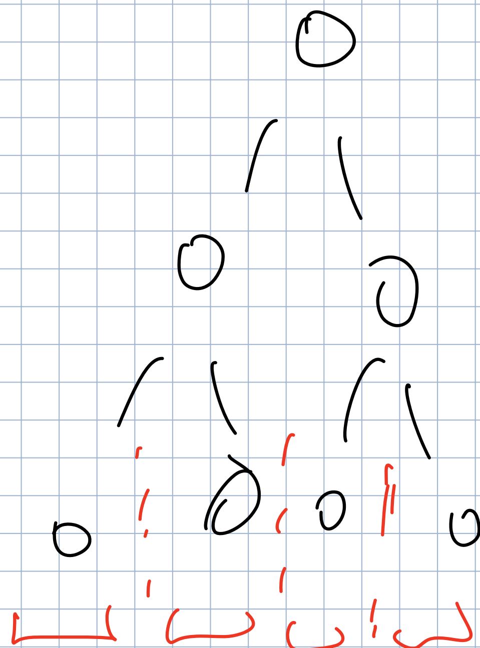
ונניח שלאחר פיצול לעומק d

נראה ש $R(d)=2^{d}$

(כלומר 2^{d} עלים)

$$R(d)=2^{d-1}\cdot 2=2^{d}$$

סה״כ יש מקסימום 2^{d} regions



log

b)

Question 5

Consider a binary decision tree used for binary classification, such as the ones we've seen in class, where each internal node splits the data based on a single feature threshold, and each leaf represents a class label. Suppose the dataset has n features, and each feature can take on real values.

(a) Prove that a binary decision tree with depth d can divide the feature space into at most 2^d regions.

(b) Explain how this result relates to the expressiveness of decision trees and their capacity to overfit.

(c) Using this result, determine the minimum depth d required for a decision tree to perfectly classify a dataset of m unique points (no two points share the same feature values).

(d) Based on the previous sections, what is the VC dimension of a tree with depth d?

(c) Using this result, determine the minimum depth d required for a decision tree to perfectly classify a dataset of m unique points (no two points share the same feature values).
(d) Based on the previous sections, what is the VC dimension of a tree with depth d?

c) בסעיף הקודם מצאנו ש-$\min d$ הוא

לכן ה-regions (העלים) הוא $2^{\min d}$

נרצה לבדוק m unique points.

כלומר m "מקרים" שונים מספיקים לנו

לכן -

$$2^{\min d} = m$$

לכן

$$\Downarrow$$

$$\min d = \log_2(m)$$

באופן כללי d (עומק) $2^d \geq m \Rightarrow \min d = \log_2(m)$

d) $VC_{dim} = 2^d$ כי מספר הנקודות המקסימלי שהעץ יכול לנפץ -

כלומר עץ עם עומק d יכול לנפץ 2^d נקודות -