שיר השקוביץ'
209144013

2 Theoretical Part

2.1 PAC Learnability

1. For $\mathcal{A}$ some learning algorithm, $\mathcal{D}$ a probability distribution over $\mathcal{X}$ and the 0-1 loss function (i.e misclassification), prove the following are equivalent:

   (a) $\forall \varepsilon, \delta > 0 \quad \exists m(\varepsilon, \delta) \quad \text{s.t} \quad \forall m \geq m(\varepsilon, \delta) \quad \mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\mathcal{A}(S)) \leq \varepsilon] \geq 1 - \delta$

   (b) $\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\mathcal{A}(S))] = 0$

   Hints:
   - For $(a) \Rightarrow (b)$ show that $\forall \varepsilon, \delta > 0$ and $\forall m \geq m(\varepsilon, \delta)$ it holds that $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \varepsilon + \delta$.
   - For $(b) \Rightarrow (a)$ use Markov's inequality

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

$\forall \varepsilon > 0 \ \exists N \in \mathbb{N}$

$\forall n > N \quad |\frac{1}{n} - 0| < \varepsilon$

$a \Rightarrow b$:

נרצה להראות שמתקיים: $\forall k>0 \quad \exists M \in \mathbb{N} \quad \forall m \geq M \quad |\mathbb{E}_{S\sim D^m}(L_D(A(S)))| < k$

יהי $k>0$. אז עבור $\varepsilon = \delta = \frac{k}{2} > 0$ נקבל $k = \varepsilon + \delta$, ומקיום (a) קיים $m(\frac{k}{2}, \frac{k}{2})$ כך שלכל $m \geq m(\frac{k}{2}, \frac{k}{2})$ מתקיים $P_{S\sim D^m}(L_D(A(S)) \leq \frac{k}{2}) \geq 1 - \frac{k}{2}$. כעת נבחין שמתקיים:

$$\mathbb{E}_{S\sim D^m}(L_D(A(S))) = E[L_D(A(S)) \leq \tfrac{k}{2}] + E[L_D(A(S)) > \tfrac{k}{2}] \leq \tfrac{k}{2} \cdot P_{S\sim D^m}(L_D(A(S)) \leq \tfrac{k}{2}) + 1 \cdot P_{S\sim D^m}(L_D(A(S)) > \tfrac{k}{2}) \leq \tfrac{k}{2} \cdot 1 + 1 \cdot \tfrac{k}{2} = k$$

$L_D(A(S)) \in [0,1]$

הסתברות תמיד קטנה שווה 1.
בנוסף מתקיים (כהסתברות משלימה):
$P_{S\sim D^m}(L_D(A(S)) > \frac{k}{2}) < \frac{k}{2}$

ומכיוון ש-k היה שרירותי הוכחנו את הנדרש.

$b \Rightarrow a$:

לפי א' שוויון מרקוב מתקיים לכל $a>0$: $P(|X| \geq a) \leq \frac{E[X]}{a}$.

נבחין ש- $L_D(A(S))$ מ"מ א' שלילי ($L_D(A(S)) \in [0,1]$). יהיו $\varepsilon, \delta > 0$.

מכך ש-(b) מתקיים, עבור $k = \varepsilon\delta > 0$ קיים $m(\varepsilon,\delta) := M \in \mathbb{N}$ כך שלכל $m \geq m(\varepsilon,\delta)$ מתקיים:

$$\mathbb{E}_{S\sim D^m}(L_D(A(S))) = |\mathbb{E}_{S\sim D^m}(L_D(A(S)))| < \varepsilon\delta$$

$L_D(A(S)) \overset{a.s}{\geq} 0$

נקבל ממרקוב:

$$P(L_D(A(S)) > \varepsilon) \leq \frac{\mathbb{E}_{S\sim D^m}(L_D(A(S)))}{\varepsilon} < \frac{\varepsilon\delta}{\varepsilon} = \delta$$

ולכן $P(L_D(A(S)) > \varepsilon) \geq 1-\delta$ כהסתברות משלימה, כנדרש.

2. Let $\mathcal{X} := \mathbb{R}^2$, $\mathcal{Y} := \{0,1\}$ and let $\mathcal{H}$ be the class of concentric circles in the plane, i.e.,

$$\mathcal{H} := \{h_r : r \in \mathbb{R}_+\} \quad \text{where} \quad h_r(\mathbf{x}) = \mathbb{1}_{[||\mathbf{x}||_2 \leq r]}$$

Prove that $\mathcal{H}$ is PAC-learnable and its sample complexity is bounded by

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \frac{\log(1/\delta)}{\varepsilon}$$

When proving, *do not* use a VC-Dimension argument. Instead, prove the claim directly from the PAC learnability definition by showing a specific algorithm and analyzing its sample complexity.
Hint: Remember that for every $\varepsilon > 0$ it holds that $1 - \varepsilon \leq e^{-\varepsilon}$

ראשית ניזכר בהגדרה ללמידות PAC:

Definition 1.1 An hypothesis class $\mathcal{H}$ is *PAC learnable* if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $\mathcal{A}$ such that:

- For every $\varepsilon, \delta \in (0,1)$
- For every distribution $\mathcal{D}$ over $\mathcal{X}$
- For every labeling function $f : \mathcal{X} \to \{0,1\}$

if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm $\mathcal{A}$ on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ *i.i.d* samples drawn from $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis $h_S = \mathcal{A}(S)$ such that:

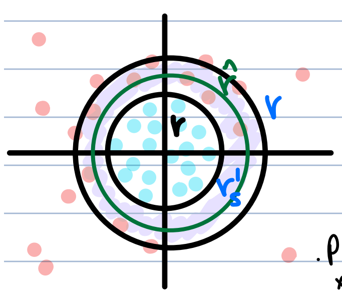
$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}(h_S) \leq \varepsilon\right] \geq 1 - \delta$$

נגדיר אלגוריתם למידה A:

בהינתן סט $S = \{(x_{1i}, x_{2i}, y_i)\}_{i=1}^{m}$ האלגוריתם A מחזיר h_{r_S} עם $\|x\| = x_{1i}^2 + x_{2i}^2$ $r_S = \max_{\substack{i: y_i = 1 \\ i \in [m]}}$.

* כך, אם נניח ריאלזביליות, כל הדוגמאות מהאימון יסווגו נכונה.

כעת נמצא את m_H (מורכבות H):



נבחין שהאזור הסגול בציור משמאל מתאר את $L_D(h_S)$. כלומר, הטעות היחידה שנוכל לקבל במבחן היא סיווג של דוגמה שאמורה להיות מסווגת 1 (כחול) מתוך 1- (כתום).
טעות כזאת תקרה כאשר מסט האימון לא מופיעה דוגמה x כך ש- $r_S < \|x\| \leq r$ (לפי האלגוריתם A).

נגדיר $\hat{r}$ רדיוס של מעגל שעבורו הסיכוי לקבל דוגמה במסגרת שבין המעגל שרדיוסו $\hat{r}$ לבין המעגל שרדיוסו r, הוא ε. כלומר- $P_{x \sim D}(x \mid \hat{r} \leq \|x\| \leq r) = \varepsilon$.

כלומר שגיאת ההכללה גדולה מ-ε אמ"מ סט האימון לא כלל אף דוגמה x שמקיימת $\hat{r} \leq \|x\| \leq r$, ולכן:

$$P_{S \sim D^m}\left(L_D(h_S) > \varepsilon\right) = P(\{S \mid \forall x \in S \ \ \|x\| \leq \hat{r} \vee \|x\| > r\}) \underset{\downarrow}{\leq} (1-\varepsilon)^m \leq \left(e^{-\varepsilon}\right)^m = e^{-\varepsilon m}$$

הדגימות מהאימון ב"ת

$$e^{-\varepsilon m} \leq \delta \iff -\varepsilon m \leq \log \delta \iff m \geq -\frac{\log \delta}{\varepsilon} \iff m \geq \frac{\log \frac{1}{\delta}}{\varepsilon}.$$

לכן אנחנו מסיקים ש-H היא למידה PAC עם $m_H(\varepsilon, \delta) = \frac{\log \frac{1}{\delta}}{\varepsilon}$.

3. Prove that if $\mathcal{H}$ has the uniform convergence property with function $m_{\mathcal{H}}^{UC}:(0,1)^2\to\mathbb{N}$, then $\mathcal{H}$ is Agnostic-PAC learnable with sample complexity $m_{\mathcal{H}}(\varepsilon,\delta)\le m^{UC}(\varepsilon/2,\delta)$.

רעיון: נעבור מהנגדרות:

Definition 4.6.4 — Uniform Convergence Property. A hypothesis class $\mathcal{H}$ is said to have the *uniform convergence property* if and only if there exists a function $m_{\mathcal{H}}^{UC}:(0,1)^2\to\mathbb{N}$ such that for every $\varepsilon,\delta\in(0,1)$ and every distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$

$$\mathcal{D}^m\left(\{S\in(\mathcal{X}\times\mathcal{Y})^m \mid S \text{ is } \varepsilon\text{-representative}\}\right)\ge 1-\delta$$

Definition 4.6.3 — ε-representative. A training sample S is called ε-representative for $\mathcal{D},\mathcal{H},\ell$ if and only if

$$\forall h\in\mathcal{H}\quad |L_S(h)-L_{\mathcal{D}}(h)|<\varepsilon$$

Definition 4.5.5 — Agnostic PAC Learnability. A hypothesis class $\mathcal{H}$ is Agnostic-PAC learnable with respect to loss $\ell:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to[0,\infty)$ if there exists a function $\tilde{m}_{\mathcal{H}}:(0,1)^2\to\mathbb{N}$ and a learning algorithm $\mathcal{A}:(\mathcal{X}\times\mathcal{Y})^m\to\mathcal{H}$ with the following property:

- For any $\varepsilon,\delta\in(0,1)$
- For any distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$

when running the learning algorithm $\mathcal{A}$ on $m\ge\tilde{m}_{\mathcal{H}}(\varepsilon,\delta)$ i.i.d samples generated by $\mathcal{D}$, the algorithm returns a hypothesis $h_S=\mathcal{A}(S)$ such that, with probability of at least $1-\delta$:

$$\mathcal{D}^m\left(\left\{S\,\middle|\,L_{\mathcal{D}}(h_S)\le\min_{h'\in\mathcal{H}}L_{\mathcal{D}}(h)+\varepsilon\right\}\right)\ge 1-\delta$$

מוגדרת אלגוריתם למידה $A\in ERM_{\mathcal{H}}$, המקיים לכל $0<\varepsilon,\delta<1$ כלשהם, מכוון של- $\mathcal{H}$ יש את התכונה של Uniform Convergence, כלומר $m_{\mathcal{H}}^{UC}(\frac{\varepsilon}{2},\delta)$ כך שמתקיים עבור $m\ge m_{\mathcal{H}}^{UC}(\varepsilon,\delta)$:

$$\mathcal{D}^m\left(\{S\in(X\times Y)^m \mid S \text{ is } \tfrac{\varepsilon}{2}\text{-representative}\}\right)\ge 1-\delta$$

רעיון שני לפי שמתקיים "את הלמה הבאה:

אזי S הוא $\frac{\varepsilon}{2}$-representative לפחות $1-\delta$ מהקלטים. אז לכל $\mathcal{H},\ell,\mathcal{D}$ כלשהם. מתקיים- $h_S\in\operatorname{argmin}_{h\in\mathcal{H}}L_S(H)$, כלומר $h_S\in ERM_{\mathcal{H}}(S)$

$$L_{\mathcal{D}}(h_S)\le\min_{h'\in\mathcal{H}}L_{\mathcal{D}}(h')+\varepsilon$$

הוכחה:

$$L_{\mathcal{D}}(h_S)\le L_S(h_S)+\frac{\varepsilon}{2}\le L_S(h)+\frac{\varepsilon}{2}\le L_{\mathcal{D}}(h)+\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=L_{\mathcal{D}}(h)+\varepsilon$$

($\frac{\varepsilon}{2}$-representative; $h_S\in ERM_{\mathcal{H}}$; $\frac{\varepsilon}{2}$-representative)

מתקיים לכל $h\in\mathcal{H}$ ולכן בהכרח את המינימום. $h=\min_{h'\in\mathcal{H}}L_{\mathcal{D}}(h')$

לכן עבור $m_H(\epsilon,\delta) := m_H(\frac{\epsilon}{2},\delta)$, נקבל לכל $m \geq m_H(\epsilon,\delta)$ שמתקיים:

$$D^m\left(\left\{S \mid L_D(h_S) \leq \min_{h' \in H} L_D(h') + \epsilon\right\}\right) \geq D^m\left(\left\{S \in (X \times Y)^m \mid S \text{ is } \frac{\epsilon}{2}\text{-representative}\right\}\right) \geq 1-\delta$$

מהלמה, ומכך שהכלת אירועים מתורגמת ל $\geq$ בהסתברויות שהאירועים קרו

2.2 VC-Dimension

1. Let $\mathcal{X} = \{0,1\}^n$ and $\mathcal{Y} = \{0,1\}$, for each $I \subseteq [n]$ define the parity function:

$$h_I(\mathbf{x}) = \left(\sum_{i \in I} x_i\right) \bmod 2.$$

What is the VC-dimension of the class $\mathcal{H}_{parity} = \{h_I \mid I \subseteq [n]\}$? Prove your answer.
Hint: what is the size of the hypothesis class?

2. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two classes for binary classification, such that $\mathcal{H}_1 \subseteq \mathcal{H}_2$. Show that $VC-dim(\mathcal{H}_1) \leq VC-dim(\mathcal{H}_2)$.

Claim 2.1 Let $\mathcal{H}$ be a finite hypothesis class then VC-Dim $(\mathcal{H}) \leq \log_2 |\mathcal{H}|$.

2. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two classes for binary classification, such that $\mathcal{H}_1 \subseteq \mathcal{H}_2$. Show that $VC-dim(\mathcal{H}_1) \leq VC-dim(\mathcal{H}_2)$.

2.3 Regularization

Based on Lecture 7 and Recitation 9

1. In the following question we will show that although the Ridge estimator is biased it can achieve lower MSE compared to the LS estimator.

   Let $\mathbf{X} \in \mathbb{R}^{m \times d}$ be a **constant** design matrix, $\mathbf{y} \in \mathbb{R}^m$ a response vector, and assume that $\mathbf{X}^\top \mathbf{X}$ is invertible. Denote $\hat{\mathbf{w}}$ the LS solution and $\hat{\mathbf{w}}_\lambda$ the ridge solution for the regularization parameter $\lambda \geq 0$ (where $\hat{\mathbf{w}}_0 \equiv \hat{\mathbf{w}}$)

   - Assume the linear model is correct, namely $\mathbf{y} = X\mathbf{w} + \varepsilon$ where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
   - Recall that in this case LS estimator is unbiased: $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}$.

   (a) Show that $\hat{\mathbf{w}}_\lambda = A_\lambda \hat{\mathbf{w}}$ where $A_\lambda := (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1}(\mathbf{X}^\top \mathbf{X})$

   (b) From the above, conclude that for any $\lambda > 0$ the ridge estimator is a biased estimator of $\mathbf{w}$. That is, show that for any $\lambda > 0$ $\mathbb{E}[\hat{\mathbf{w}}_\lambda] \neq \mathbf{w}$.

   (c) Show that: $\mathrm{Var}(\hat{\mathbf{w}}_\lambda) = \sigma^2 A_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} A_\lambda^\top$, for σ^2 the variance of the assumed noise.
   Hint:: Recall that for a constant matrix B and a random vector $\mathbf{z}$ it holds that $\mathrm{Var}(B\mathbf{z}) = B \cdot \mathrm{Var}(\mathbf{z}) \cdot B^\top$ and that $\mathrm{Var}(\hat{\mathbf{w}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

   (d) Derive explicit expressions for the (squared) bias and variance of $\hat{\mathbf{w}}_\lambda$ as a function of λ, i.e. write a bias-variance decomposition for the mean square error of $\hat{\mathbf{w}}_\lambda$.

   Hint: recall that for the multivariate case the MSE defined to be:

   $$MSE(\hat{\mathbf{y}}) = \mathbb{E}\left[\|\hat{\mathbf{y}} - \mathbf{y}\|^2\right] = \mathbb{E}\left[(\hat{\mathbf{y}} - \mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y})\right]$$

   where $\mathbf{y}$ is the true value, and $\hat{\mathbf{y}}$ is our estimation.

   (e) The derivative of the MSE with respect to λ is negative at $\lambda = 0$. We omit here the proof. Conclude that, if the linear model is correct, a little Ridge regularization helps to reduce the MSE.

(a)

$$A_\lambda \hat{\omega} = (X^TX + \lambda I_d)^{-1}(X^TX) \cdot \hat{\omega} = (X^TX + \lambda I_d)^{-1} \underbrace{(X^TX)(X^TX)^{-1}}_{I} X^T y = (X^TX + \lambda I_d)^{-1} X^T y = \hat{\omega}_{ridge}$$

$\hat{\omega} = (X^TX)^{-1} X^T y$ מפיתוחים שראינו ל- LS ומכך ש- X^TX הפיכה

ראינו בתרגול שמתקבל כתוצאה מגזירה והשוואה ל-0: $\hat{\mathbf{w}}_\lambda^{ridge} = (\mathbf{X}^\top \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^\top \mathbf{y}$

(b)

$$E[\hat{\omega}_\lambda] = E[A_\lambda \hat{\omega}] = A_\lambda E[\hat{\omega}] = A_\lambda \omega$$

λ, X קבועים ולכן A_λ קבועה

$\hat{\omega}$ אומד לא מוטה

נבחין שעבור $\lambda>0$ $A_\lambda \neq I$ כי $\lambda=0 \leftrightarrow A_\lambda = (X^TX)^{-1}(X^TX) = I$. לכן $w \neq A_\lambda w$ עבור $\lambda>0$
ולכן מצאנו שהאומד בממוצע שונה מהפרמטר האמיתי ולכן האומד מוטה.

(c)

סעיף א'

$\hat{\omega}$ וקטור מקרי, A_λ קבועה

$$\mathrm{Var}(\hat{\omega}_\lambda) = \mathrm{Var}(A_\lambda \hat{\omega}) = A_\lambda \mathrm{var}(\hat{\omega}) A_\lambda^T = A_\lambda \sigma^2 (X^TX)^{-1} A_\lambda^T = \sigma^2 A_\lambda (X^TX)^{-1} A_\lambda^T$$

σ^2 סקלר ולכן ניתן להזיזו

(d) Derive explicit expressions for the (squared) bias and variance of $\hat{\mathbf{w}}_\lambda$ as a function of λ, i.e. write a bias-variance decomposition for the mean square error of $\hat{\mathbf{w}}_\lambda$.

Hint: recall that for the multivariate case the MSE defined to be:

$$MSE(\hat{\mathbf{y}}) = \mathbb{E}\left[\|\hat{\mathbf{y}} - \mathbf{y}\|^2\right] = \mathbb{E}\left[(\hat{\mathbf{y}} - \mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y})\right]$$

(d) ראינו בתרגול עבור המקרה החד מימדי שמתקיים:

$$\begin{aligned}
\underset{MSE}{\mathbb{E}\left[(\hat{\mathbf{y}} - \mathbf{y})^2\right]} &= \mathbb{E}\left[[(\hat{\mathbf{y}} - \bar{\mathbf{y}}) + (\bar{\mathbf{y}} - \mathbf{y})]^2\right] \\
&= \mathbb{E}\left[(\hat{\mathbf{y}} - \bar{\mathbf{y}})^2\right] + 2(\bar{\mathbf{y}} - \mathbf{y})\mathbb{E}[\hat{\mathbf{y}} - \bar{\mathbf{y}}] + (\bar{\mathbf{y}} - \mathbf{y})^2 \\
&= \mathbb{E}\left[(\hat{\mathbf{y}} - \bar{\mathbf{y}})^2\right] + (\bar{\mathbf{y}} - \mathbf{y})^2 \\
&= \text{var}[\hat{\mathbf{y}}] + \text{bias}^2[\hat{\mathbf{y}}]
\end{aligned}$$

כש $\bar{y} = E[\hat{y}]$

In words,

$$\text{MSE} = \text{Variance} + \text{bias}^2$$

נרחיב למקרה הרב מימדי ונקבל באופן דומה:

$$MSE = E\left(\|(\hat{\omega}_\lambda - E(\hat{\omega}_\lambda)) + (E(\hat{\omega}_\lambda) - \omega)\|^2\right) =$$

לינאריות התוחלת, $\|u-v\|^2 = \sum_i (u_i - v_i)^2 = \sum_i u_i^2 - 2u_iv_i + v_i^2 = \|u\|^2 - 2\langle u,v\rangle + \|v\|^2$

$$= E(\|\hat{\omega}_\lambda - E(\hat{\omega}_\lambda)\|^2) + 2(E(\hat{\omega}_\lambda) - \omega)^t E(\hat{\omega}_\lambda - E(\hat{\omega}_\lambda)) + E(\|E(\hat{\omega}_\lambda) - \omega\|^2) =$$

$$= E(\|\hat{\omega}_\lambda - E(\hat{\omega}_\lambda)\|^2) + \|E(\hat{\omega}_\lambda) - \omega\|^2 = var_\omega(\lambda) + Bias^2_\omega(\lambda)$$

$E(\hat{\omega}_\lambda - E(\hat{\omega}_\lambda)) = E(\hat{\omega}_\lambda) - E(\hat{\omega}_\lambda) = 0$

$\omega, E(\hat{\omega}_\lambda)$ קבועים

השוויון האחרון נכון עבור משערך רב מימדי, וניתן להמשיך ולפתח ולמצוא ש:

$$Bias^2_\omega(\lambda) = (E(\hat{\omega}_\lambda) - \omega)^t (E(\hat{\omega}_\lambda) - \omega) = \|E(\hat{\omega}_\lambda) - \omega\|^2 = \|A_\lambda \omega - \omega\|^2 = \|(A_\lambda - I)\omega\|^2$$

w רב מימדי ; סעיף b

$$Var_\omega(\lambda) = E(\|\hat{\omega}_\lambda - E(\hat{\omega}_\lambda)\|^2) = E\left(\sum_i (\hat{\omega}_{\lambda_i} - E(\hat{\omega}_\lambda)_i)^2\right) = \sum_i E(\hat{\omega}_{\lambda_i} - E(\hat{\omega}_\lambda)_i)^2 =$$

w רב מימדי ; מסעיף א', ב'

$$\sum_i var(\hat{\omega}_{\lambda_i}) = Trace(var(\hat{\omega}_\lambda)) = Trace(\sigma^2 A_\lambda (X^TX)^{-1} A_\lambda^T) = \sigma^2 Trace(A_\lambda (X^TX)^{-1} A_\lambda^T)$$

סעיף ג'

לכן סה"כ קיבלנו ש:

$$MSE(\hat{\omega}_\lambda) = var(\lambda) + Bias^2_\omega(\lambda) = \sigma^2 Trace(A_\lambda (X^TX)^{-1} A_\lambda^T) + \|(A_\lambda - I)\omega\|^2$$

(e) The derivative of the MSE with respect to λ is negative at $\lambda = 0$. We omit here the proof. Conclude that, if the linear model is correct, a little Ridge regularization helps to reduce the MSE.

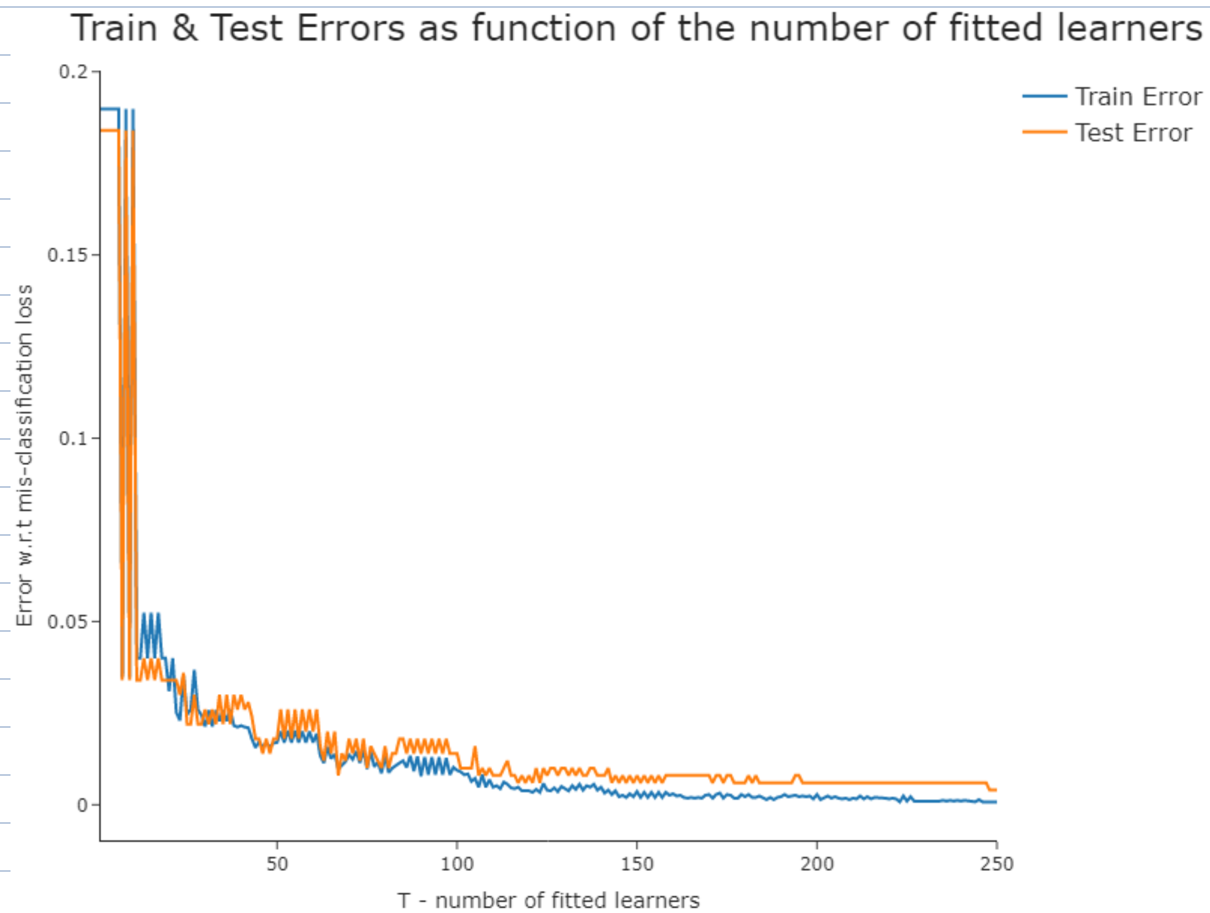
לאורך הסעיפים הנחנו שהמודל הלינארי נכון ובהתבסס על כך הגענו למסקנה כיצד ה- MSE ניתנת לכתיבה כפירוק של ההטיה והשונות של המשערך $\hat{w}_\lambda$. מהנתון, ניתן להראות שכאשר נגזור ביטוי זה עבור $\lambda=0$ נקבל נגזרת שלילית. אם הנגזרת של ה- MSE כש- $\lambda=0$ שלילית, מהגדרת הנגזרת המשמעות היא שהגדלה של λ מקטינה את הפונקציה, ולכן שימוש בגורם רגולריזציה $0<\lambda$ קטן, כלומר שימוש במשערך $\hat{w}_\lambda$ עם $0<\lambda$ קטנה מספיק, במקום שימוש ב- $\hat{w}=w_0$, תקטין את השגיאה, את ה- MSE.

3 Practical Part

3.1 Boosting - Separate the inseparable

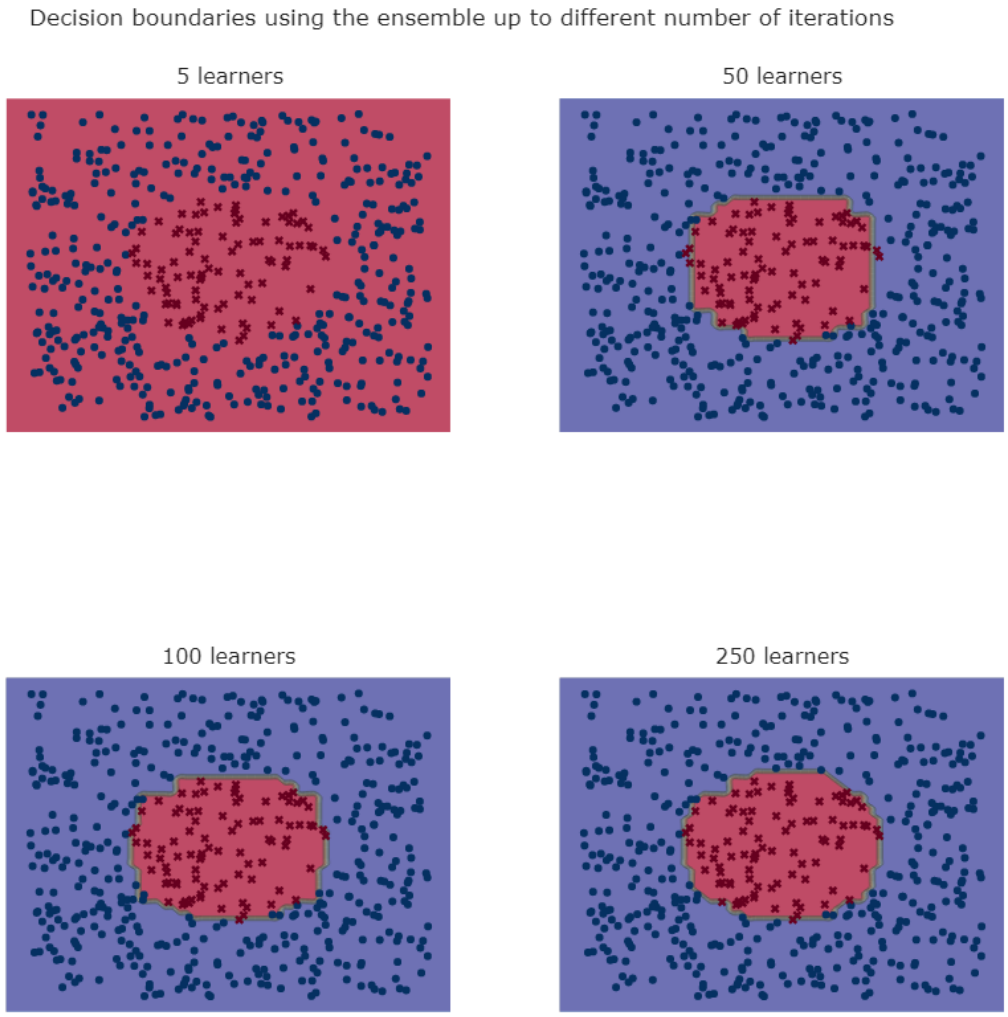
1. Use the provided `generate_data` function to generate 5000 train samples and 500 test samples, both with no noise (i.e. `noise_ratio = 0`). Train an Adaboost ensemble of size 250 (i.e passing 250 as the number of iterations) using your implementation of the `DecisionStump` as weak learner. Plot, in a single figure, the training- and test errors as a function of the number of fitted learners (i.e. use the `partial_loss` function for $t = 1, \ldots, 250$). Explain your results.

.1



אנחנו רואים שהגדלת מספר הלומדים החלשים (במקרה זה עצי החלטה מעומק 1) בהם משתמש מסווג Adaptive Boosting בכדי לקבל החלטה כיצד לסווג כל דגימה, הן על פני סט האימון והן על פני סט המבחן, מצליחה להקטין את השגיאה, כלומר, להביא לסיווג טוב יותר של הסט. נבחין שהביצוע על פני סט המבחן מובלה כמעט כמו הביצוע על סט האימון, כלומר כמעט ואין overfit לדאטה שבאימון. תוצאות אלה עומדות בקנה אחד עם מה שלמדנו בכיתה ובתרגול - adaboost משתמש בלומד נוסף בכל איטרציה בניסיון לתייג נכון דוגמאות שהלומדים הקודמים לא הצליחו לתייג נכון, וממשקל כל לומד לפי מידת הצלחתו בתיוג. כך, הלומד משקלל לומדים רבים שהותאמו להתפלגויות שונות ובכך מצליח לסווג נכונה יותר דוגמאות וכן תורם למניעה של overfit (כי מנסים לסווג נכון את כל הדגימות ולאו דווקא דגימות שכיחות בסט האימון שאולי לא משקפות את ההתפלגות האמיתית של הדאטה). בנוסף, הדיוני שהשגיאה של adaboost מתייצבת בשלב כלשהו ועד מגיעה בדיוק ל-0 גם כשמוסיפים עוד לומדים, הן עבור סט האימון והן עבור סט המבחן. בנוסף נבחין שהעלאה של מספר הלומדים מ-0 עד בערך 50 מביאה לשיפור משמעותי, ולאחר מכאן השינוי ברמת הדיוק קטן יותר עם הוספת לומדים נוספים. בערך מ-100 איטרציות המסווג מבצע סיווג כמעט מושלם הן על סט המבחן והן על סט האימון.

2. Using the previously fitted ensemble, plot the decision boundary obtained by using the the ensemble up to iteration 5, 50, 100 and 250. Use your implementation of the `partial_predict` to obtain the predictions of the ensemble up to the specified size. In each of these plots also add the test set (colored and/or shaped by the actual labels). Explain your results. .2



ניתן לראות מהצבעים של הנקודות שמשקפים את הלייבלים האמיתיים שהדאטה מורכב ממעגל במרכז ששייך לסיווג אחד, ונקודות שמחוץ למעגל משויכות לסיווג אחר. על כן, זהו מידע שלא ניתן להפריד באופן לינארי. על כן לומד חלש יחיד לא יצליח לסווג נכונה את המידע.

אך אנחנו רואים שככל שיש יותר לומדים שמסווג ה- ensemble (מודל ה-adaboost) רשאי להשתמש בהם, כך הוא מצליח לסווג באופן מדויק יותר ויותר דוגמאות, ולייצר קווי החלטה מורכבים יותר מאשר מודלים לינאריים (עצי החלטה בעומק 1 מסוגלים לבצע הפרדה לינארית בלבד).

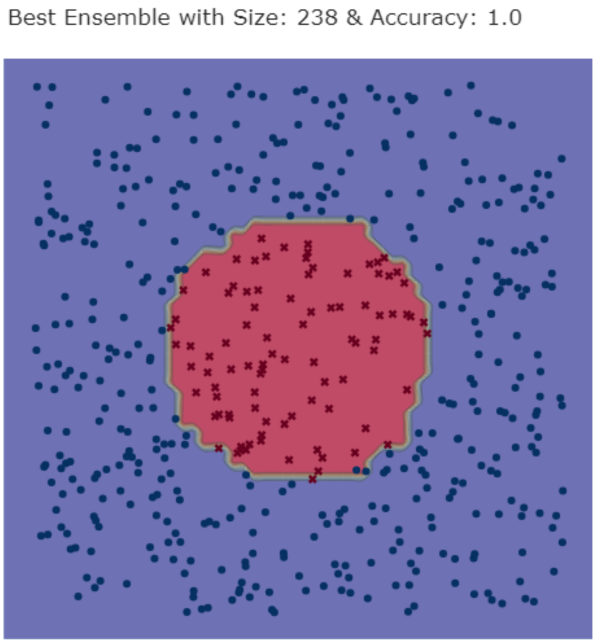
כך למשל עבור 250 איטרציות ה-adaboost כבר מצליח לייצר קו החלטה מעגלי יחסית שתואם לדאטה.

התוצאות בסעיף זה עומדות בקנה אחד עם התוצאות מהסעיף הקודם- ככל שיש יותר איטרציות, יש יותר לומדים שה-ensemble משתמש בהם, השגיאה קטנה, ואכן אנחנו רואים מדוע היא קטנה- מכיוון שקווי ההחלטה שמצליחים לייצר הם מורכבים יותר ויותר מותאמים לדאטה.

כמו כן ניתן לראות מדוע בהתחלה עם הוספת לומדים יש צניחה משמעותית בשגיאה ולאחר מכן הצניחה פחות משמעותית- בהתחלה מצליחים לייצר קווי החלטה לא לינאריים שיחסית מייצגים נכונה את הדאטה (המעבר מ-5 לומדים ל-50), ולאחר מכן נשארה מסה יחסית קטנה של דגימות "קשות לסיווג", שדורשות הבנה מדויקת של הדאטה, ולכן נדרשת העלאה יחסית גדולה של מספר הלומדים כדי לקבל קווי החלטה מורכבים שיצליחו לסווג נכון, וכשהם מצליחים לייצר, מעט הדגימות הקשות מצליחות לקבל סיווג נכון ולהוריד במעט את השגיאה שנותרה.

3. Using the test evaluation scores per number of learners, which ensemble size achieved the lowest test error? Plot the decision surface of that ensemble as well as the test set data points. In the plot's title provide the ensemble size and its accuracy.

.3

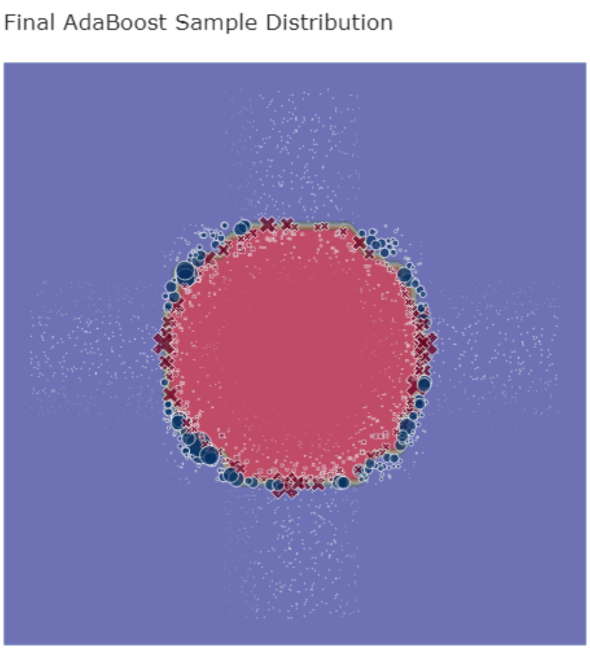


המודל שמשיג את שגיאות המבחן המינימלית (במקרה זה 0 עם עיגול של שתי נקודות אחרי הנקודה) הוא זה עם ההיפר פרמטר של מס' איטרציות 238.

4. Using the previously fitted ensemble, use the weights of the last iteration (i.e. D at T) to plot the *training* set with a point size proportional to its weight and color (and/or shape) indicating its label.
   - As previously, plot the decision surface (using the full ensemble).
   - As the weights are of very small numbers normalize and transform as follow: `D = D/np.max(D) * 5`

   Explain your results, and specifically explain which samples are "easier" and which are more "challenging" for the classifier.

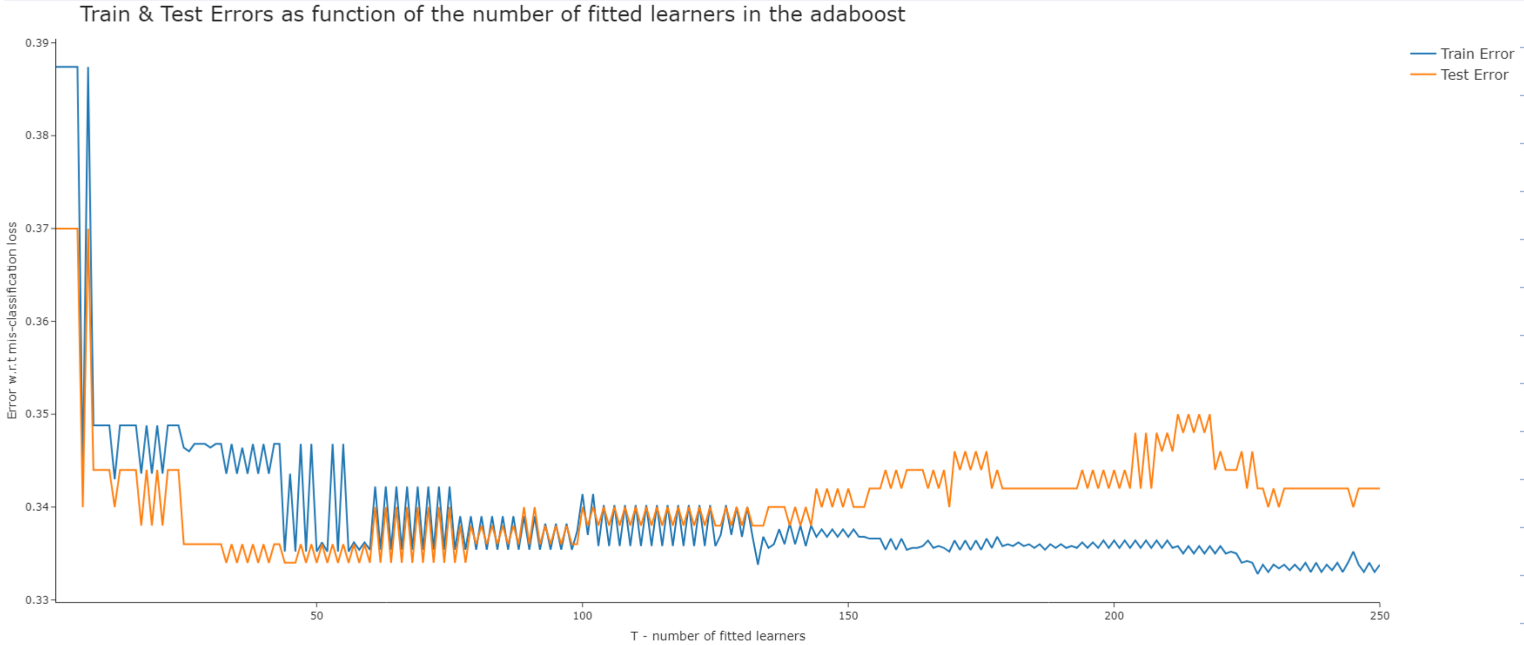
.4



אנחנו רואים שהדגמאות שההסתברות לדגום אותן באיטרציה האחרונה גדולה במיוחד, אלו דוגמאות שיושבות באזור היקף המעגל שמתאר את קו ההחלטה הנדרש עבור הפרדה. המודל האופטימלי שמצאנו, כלומר המודל באיטרציה האחרונה, נותן הכי הרבה הסתברות לדוגמאות שהוא עד לא מסווג טוב, כי הן דורשות גבול מסובך, היקף מעגל חלק ומורכב שקשה להשיג עם לומדים לינאריים. לעומת זאת דגימות שיושבות יחסית רחוק מהיקף המעגל בעלות הסתברות נמוכה להידגם באיטרציה האחרונה, עדות מיזון שהמודל כבר הצליח למצוא לומדים שתייגו אותן נכון, הן פחות מאתגרות.

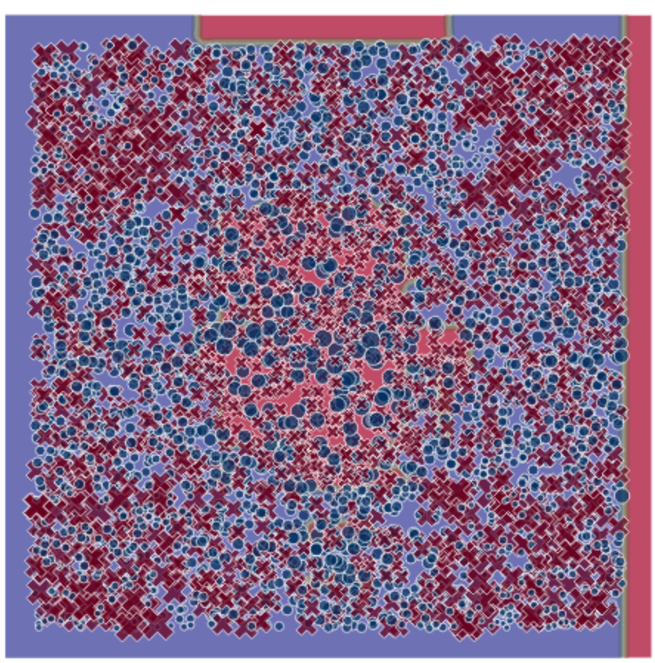
5. Repeat the steps above (while avoiding code repetition) for train- and test sets generated with noise levels of 0.4. Show graphs as in question (1) and question (4). Explain the results. In your answer explain what is seen in the plot of the loss in terms of the bias-variance tradeoff.

.5



ניתן ללמוד מהגרף, שכאשר מכניסים רעש לדאטה, מקבלים התנהגות שונה של ה-adaboost. כעת, בחירה של הרבה לומדים (בערך מעל 150) מגדילה את ה-overfit, מה שמתבטא בשגיאת מבחן שהולכת וגדלה ושגיאת אימון שהולכת וקטנה. מאופן פעולת adaboost, זוהי תוצאה הגיונית - המודל מנסה לסווג נכון את כל הדוגמאות, גם outliers ומידע רועש, מאחר שהוא מנסה לסווג נכונה גם דגימות קשות. אם מניחים שהדגימות הללו הן רעש, לנסות לסווג נכון משמע overfit. כלומר מבחינת bias-variance tradeoff, לקיחת הרבה לומדים, מס' איטרציות גבוה, עבור ה-adaboost, על מידע רועש, עשוי להוביל ל- variance גבוה שנובע מניסיון לרדוף אחר הרעש, אך מצד שני מאפשר לייצר קווי החלטה מורכבים יותר, ולכן ה- bias נמוך יותר.

Final AdaBoost Sample Distribution



ואכן, במקרה זה הלומד הטוב ביותר זה הוא עם מספר נמוך של לומדים, עם 33 איטרציות (למניעת overfit).

נבחין שמתקבל Accuracy נמוך בהרבה מהמקרה ללא הרעש (~0.7) וזה מתבטא גם בכך שיש הרבה דגימות שבאיטרציה האחרונה ההסתברות לדגום אותן גבוהה (במודל האופטימלי). זה מכיוון שכעת יש הרבה דגימות שמאתגרות את המודל, וקשה יותר להתחקות אחר קווי החלטה שמביאים לסיווג נכון את כל הדגימות.

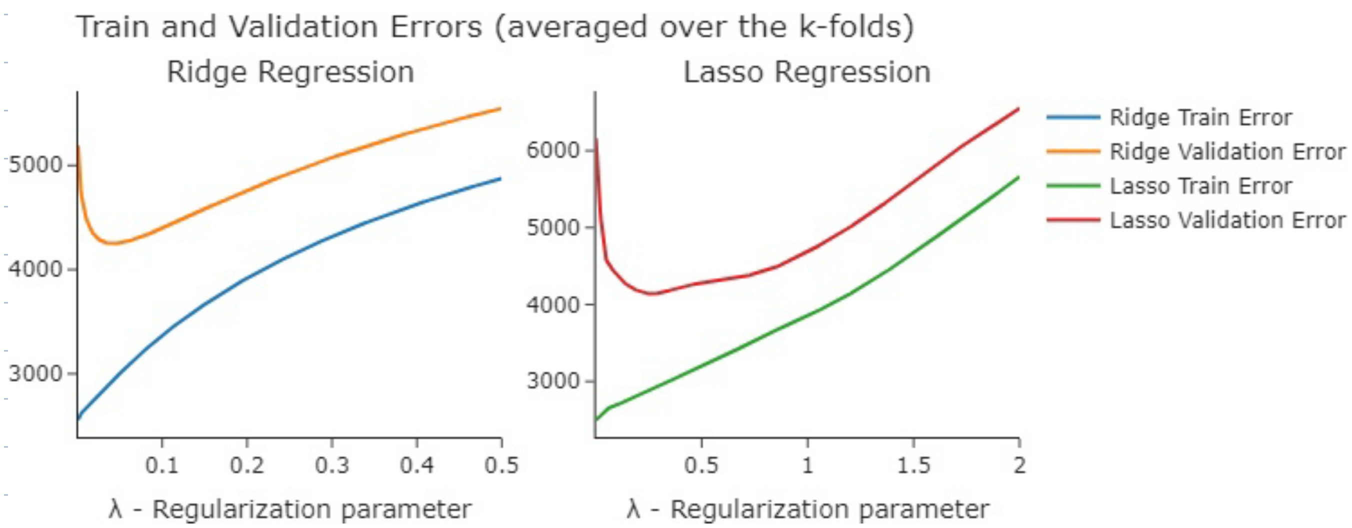
* וניתן לראות שהדוגמאות שבהן סווגו נכון עם הסתברות נמוכה יותר.

3.2 Choosing Regularization Parameters Using Cross Validation

2. For both Ridge (with intercept) and Lasso regularizations, run 5-Fold Cross-Validation using different values of the regularization parameter λ. Explore possible ranges of values of λ (say, take `n_evaluations=500` equally spaced values in a range of your choice). Which range of values is meaningful for the given data for each of the algorithms?

   Then, over these selected ranges, for each of the two algorithms, plot the train- and validation errors as a function of the tested regularization parameter value. Explain your results. Address both the differences between the train- and validation errors as well as the differences in the plots for the two algorithms.

התחום המשמעותי עבור Ridge: [0,0.5]
התחום המשמעותי עבור lasso: [0,2]
אלו האזורים המשמעותיים מכיוון ששם משיגים מינימום שגיאה על הטסט.



ניתן לראות בשני הגרפים שהשגיאה על פני ה- validation set גבוהה מאשר על ה- training set. אנחנו רואים שעבור ערכים מסוימים של ההיפר פרמטר יש יותר הבדל מאשר באחרים, ושמתקבלת שגיאה קטנה יותר בסט ה- validation מאשר במקרים אחרים (כאשר זה משקף את ההצלחה המשוערת ב- test). לכן נבחר את ההיפר פרמטר שנותן את השגיאה הקטנה יותר על פני ה- validation set.

* השגיאות הן convex ולכן מינימום מקומי הוא גלובלי.

עבור סט האימון בשני המקרים המינימום מושג ב-0, כלומר במצב בו אין פקטור של רגולריזציה. שזוהי תוצאה הגיונית משום שבמצב זה המטרה היחידה היא מזעור שגיאת האימון ואין ניסיון לבחור מודל פשוט על חשבון הגדלת שגיאה זאת. לעומת זאת, המינימום עבור שגיאות ה-validation שמשקפת את שגיאות ההכללה, מושגת כאשר ישנו פקטור רגולטורי, תוצאה שמתאימה לתערו של אוקאם, לפיו עדיף לבחור במודל הפשוט ביותר שמסביר טוב מספיק את הדאטה, ויותר סביר להניח ולכן לפיו גם אם שגיאות האימון לא שווה 0, כשמתחשבים בשגיאות ההכללה עדיף לבחור מודל פשוט ככל הניתן.

בנוסף נשים לב שעבור המודלים השונים מתאים היפר פרמטר שונה:
זה הגיוני מאחר ש- Ridge משתמש ב- l_2 ו- lasso משתמש ב- l_1 ולכן סביר שההיפר פרמטרים יהיו שונים.

3. Report which regularization parameter values achieved the best validation errors for the Ridge and Lasso regularizations. Then fit the entire test set using these values for Ridge, Lasso and Least Squares regressions. Report the test errors of each of the fitted models.
    - Use your *own* previously implemented Least Squares algorithm.

```
Best hyperparameter (regularization values):
 Ridge Regression: 0.042
 Lasso Regression: 0.2613907815631263
```

```
Mean Square Error for different Regression Models:
 Ridge Regression: 3128.948548647699
 Lasso Regression: 3188.714771629627
 Least Squares Regression: 3425.32159754408
```

(over test set)