מטריצת אימון: $X\in\mathbb{R}^{d\times m}$ ששורותיה הן תכונות, ועמודותיה הן הדגימות. **וקטור תיוג:** $y\in\mathbb{R}^m$ בעל ערך $y_i=f(x_i)+z_i$ עבור z_i רעש כלשהו.

Empirical Risk

Empirical Risk: עבור פונקציית מחיר $\ell:\mathbb{R}^d\to\mathbb{R}$:
$$ER(h)=\frac{1}{m}\sum_{i=1}^m \ell(h(x_i),y_i)$$

Squared Loss (RSS)	$\sum_{i=1}^m (h(x_i)-y_i)^2=\|h(X)-y\|_2^2$
Absolute Value Loss	$\sum_{i=1}^m\lvert h(x_i)-y_i\rvert=\|h(X)-y\|_1$
0-1 Loss	$\frac1m\sum_{i=1}^m 1_{[h(x)\neq y]}$

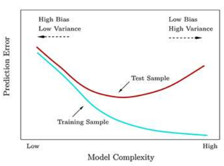
Maximum Likelihood

Likelihood Function: בהינתן התפלגות $\mathcal{D}(\Theta)$ עם פרמטר Θ, ודגימות $x_1,\dots,x_m\overset{iid}{\sim}\mathcal{D}(\Theta)$ אנו מגדירים את הנראות:
$$L(\Theta)=L(\Theta|x_1,\dots,x_m)=\Pr_{\mathcal{D}}[x_1,\dots,x_m|\Theta]$$
MLE (ר"ת Maximum Likelihood Estimator): יהי $X=x_1,\dots,x_m\overset{iid}{\sim}\mathcal{D}(\Theta)$ קבוצת דגימות, אז הנראות המרבית:
$$\widehat{\Theta}_{MLE}=\arg\max_\Theta L(\Theta|X)$$

Bias & Variance

Bias: הסטייה של המודל מנקודות המדגם (שגיאת ההכללה). ככל שקבוצת ההיפותזות גדולה יותר, כך הסטייה קטנה. נובעת מ-Approximation Error.

Variance: השונות של המודל מהאמת (מרדף אחרי הרעש). ככל שקבוצת ההיפותזות גדולה יותר, כך השונות גדלה. נובעת מ-Estimation Error.



רגרסיה לינארית

$$\mathcal{H}_{reg}=\left\{h(x)\middle|h(x)=\left\langle w,\binom{1}{x}\right\rangle, w\in\mathbb{R}^{d+1}\right\}$$
(w_0 נקרא intercept. $w_1,\dots,w_d$ נקראים המשקולות)

Normal Equations: מזעור ה-RSS שקול ל- $X(y-X^Tw)=0$.

טענה: התאים הבאים שקולים: **1.** קיים פתרון יחיד למערכת. **2.** $\dim(Ker(X^T))=0$. **3.** XX^T הפיכה. **4.** $\sigma_{d+1}(X)>0$.

פתרון: $\hat w=(XX^T)^{-1}Xy$. (בעל נורמה מינימלית).

טענות מההוכחה: 1. $Ker(X^T)=Ker(XX^T)$. **2.** עבור מטריצה A ריבועית, $Im(A^T)=Ker(A)^\perp$. **3.** תהי $y=X^Tw$ מערכת לא הומוגנית, ונניח X^T לא הפיך. למערכת אינסוף פתרונות אמ"ם $y\perp Ker(X)$. **4.** למערכת $XX^Tw=Xy$ יש אינסוף פתרונות או פתרון יחיד (אם XX^T הפיכה).

משפט: $\hat w=(X^T)^\dagger y$ הוא תמיד פתרון למשוואה הנורמלית.

הגדרה: אם X^T כולל עמודה (פיטצ'ר) הנפרש ע"י יתר העמודות נאמר שהוא כמעט סינגולרי (=לא הפיך). במקרה זה ניקח פתרון:
$$X^{\dagger,\epsilon}=V\Sigma^{\dagger,\epsilon}U^T\ \ s.t\ [\Sigma^{\dagger,\epsilon}]_{ii}=\begin{cases}\frac{1}{\sigma_i} & \sigma_i>\epsilon\\ 0 & otherwise\end{cases}$$

MLE: עבור $z_i\overset{iid}{\sim}N(0,\sigma^2)$ (כלומר $y_i\overset{iid}{\sim}N(x_i^Tw,\sigma^2)$) אז הצפיפות:
$$p(y|w)=\prod_{i=1}^m\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(x_i^Tw-y_i)^2}{2\sigma^2}}$$
$$=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^m}\prod_{i=1}^m e^{-\frac{(x_i^Tw-y_i)^2}{2\sigma^2}}=L(w|y)$$
נראות מרבית:
$$\hat w=\arg\max_w L(w|y)\overset{by\ log}{=}\arg\min_w\sum_{i=1}^m(x_i^Tw-y_i)^2$$

Bias-Variance Tradeoff: מתקיים-
$$MSE=\mathbb{E}[(\hat y-y)^2]=Var[\hat y]+bias^2[\hat y]=Variance+bias^2$$

קלסיפיקציה

סיווג שגיאות: שגיאת **Type-I** (או FP) זו השגיאה החמורה. שגיאת **Type-II** (או FN).

	Positive	Negative	
True	TP, Recall, TPR: $\frac{TP}{Positive}$	TN, Specificity: $\frac{TN}{Negative}$	Accuracy: $\frac{TP+TN}{Total}$
False	FP, FPR: $\frac{FP}{Negative}$	FN, FNR: $\frac{FN}{Positive}$	Error Rate: $\frac{FP+FN}{Total}$
	Precision: $\frac{TP}{TP+FP}$	Negative predictive value: $\frac{TN}{TN+FN}$	$F_1=2\cdot\frac{Recall\cdot Precision}{Recall+Precision}$

Half-spaces

$$\mathcal{H}_{half}=\{x\mapsto sign(\langle x,w\rangle+b)\}$$
תחת ההנחה שקיים מפריד לינארי, נגדיר את הסיכון להיות $L_S(h_w)=\sum_{i=1}^m 1_{\{y_i\cdot\langle x,w\rangle<0\}}$ ונמזער לפי עקרון ה-ERM.
$$\arg\min_w 0 \quad s.t\ \ diag(y)\cdot X^Tw\ge 1$$

Perceptron Algorithm:

Initialize: $w=\vec{0}$

While $\exists i$ s.t $y_i\cdot\langle x_i,w\rangle\le0$:

$w\leftarrow w+y_i\cdot x_i$

SVM

$$\mathcal{H}_{SVM}=\{x\mapsto sign(\langle x,w\rangle+b)\}$$
מרחק: $D\left(\binom{w}{b},x_i\right)=\min_{v:\langle v,w\rangle+b=0}\|x_i-v\|$

שוליים: $M\left(\binom{w}{b},S\right)=\min_{x_i\in S}D\left(\binom{w}{b},x_i\right)$

$$Hard-SVM=\arg\min_{w,b}\|w\|^2\quad s.t\ \ y_i\cdot(\langle x_i,w\rangle+b)\ge1\qquad \left(margin=\frac{2}{\|w\|}\right)$$
$$Soft-SVM=\arg\min_{w,b}\lambda\cdot\|w\|^2+\frac1m\sum_{i=1}^m\xi_i\quad s.t\ \ y_i\cdot(\langle x_i,w\rangle+b)\ge1-\xi_i,\ \xi_i\ge0$$
$$\Leftrightarrow\arg\min_w\lambda\cdot\|w\|^2+\frac1m\sum_{i=1}^m\underbrace{\max\{1-y_i\cdot(\langle x_i,w\rangle+b),0\}}_{\ell^{hinge}(y_i\cdot(\langle x_i,w\rangle+b))}$$

Kernels: מציאת העתקה $\psi:\mathcal{X}\to F$ (באשר F מרחב מכפלה פנימית שלם).
$$w^*=\arg\min_w\lambda\cdot\|w\|^2+f(\langle w,\psi(x_1)\rangle,\dots,\langle w,\psi(x_m)\rangle)$$
טענה: קיים $\alpha\in\mathbb{R}^m$ עבורו $w^*=\sum_i\alpha_i\psi(x_i)$. במקרה זה נסמן $K(x_i,x)=\langle\psi(x_i),\psi(x)\rangle$ ואז:
$$\langle w^*,\psi(x)\rangle=\sum_{i=1}^m\alpha_iK(x_i,x)$$

Bayes Optimal Classifier

בהינתן התפלגות $\mathcal{D}$ מעל $\mathcal{X}\times\mathcal{Y}$, עבור הפסד $L_{\mathcal{D}}(h)=\mathbb{E}_{(x,y)\sim\mathcal{D}}[h(x)\neq y]$, נגדיר:
$$f_{\mathcal{D}}=\arg\max_{y\in\mathcal{Y}}\Pr_{\mathcal{D}}[Y=y|X=x]$$
טענה: עבור היפותזה $h:\mathcal{X}\to\mathcal{Y}$ מתקיים $L_{\mathcal{D}}(f_{\mathcal{D}})\le L_{\mathcal{D}}(h)$.

LDA: עבור $\mathcal{X}=\mathbb{R}^d$ והתפלגות $x|y\sim\mathcal{N}(\mu_y,\Sigma)$, כלומר:
$$f(x|y)=\frac{1}{\sqrt{(2\pi)^d\det(\Sigma)}}e^{-\frac12(x-\mu_y)^T\Sigma^{-1}(x-\mu_y)}$$
מתקיים:
$$h_{\mathcal{D}}(x)=\arg\max_y\underbrace{x^T\Sigma^{-1}\mu_y-\frac12\mu_y^T\Sigma^{-1}\mu_y+\ln(\Pr[y])}_{\delta_y(x)}$$

Logistic Regression

בהינתן $y=Xw+z$ (עבור $z_i\overset{iid}{\sim}N(0,\sigma^2)$), פונקציה $\phi:\mathbb{R}\to[0,1]$ (מונוטונית עולה, חלקה, חח"ע מהישר הממשי לקטע $(0,1)$). נניח $y_i\sim Ber(p_i)$ כאשר $p_i=\phi(\langle x_i,w\rangle)$ עבור $w\in\mathbb{R}^{d+1}$.

נשתמש בפונקציה הלוגיסטית $\pi(x)=\frac{e^x}{1+e^x}=\frac{1}{1+e^{-x}}$:
$$\mathcal{H}^d_{logistic}=\{x\mapsto\pi(\langle x,w\rangle)\}$$
התועלת $L(w|y)=\prod p_i(w)^{y_i}(1-p_i(w))^{1-y_i}$, ובעזרת log:
$$\hat w=\arg\max_{w\in\mathbb{R}^{d+1}}\sum_{i=1}^m\left[y_i\langle x_i,w\rangle-\log\left(1+e^{\langle x_i,w\rangle}\right)\right]$$
מציאת פתרון ע"י לקיחת מינוס ושימוש באיטרציות ניוטון-רפסון.

חיתוך (cutoff): בגלל שאנו מעוניינים בתוצאה מתוך $\mathcal{Y}$ נבצע-
$$\hat y=\begin{cases}1 & h_w(x)\ge\alpha\\0 & h_w(x)<\alpha\end{cases}$$
עקומת ROC (ר"ת Receiver Operating Characteristic) מתאימה בין TPR לבין FPR. אנו משתמשים **במדד AUC** (ר"ת Area Under the Curve) בשביל למצוא α עבור w נתון.

צורה נוספת ל-Logistic Regression: (זו פונקציה קמורה!!)
$$\arg\min_w\frac1m\sum_{i=1}^m\log\left(1+e^{-y_i\langle w,x_i\rangle}\right)$$

Nearest Neighbors

נגדיר מרחק ממושקל $\rho(x_i,z)=\sum_{j=1}^d w_j\cdot(x_j-y_j)^2$, ועבור k השכנים הקרובים נבחר:
$$\arg\max_{y\in\mathcal{Y}}\sum_{i=1}^k 1_{\{y_{\pi(i)}=y\}}$$
מציאת השכנים ע"י Brute Force, עיבוד מקדים (למשל kd-tree) או חיפוש רנדומלי מהיר. (אין עקרון למידה).

Classification Trees

$$\mathcal{H}^k_{CT}=\left\{x\mapsto\sum_{j=1}^N c_j\cdot 1_{\{x_i\in B_j\}}\right\}$$
N כמות הקופסאות, k עומק מקסימלי. $\mathbb{R}^d=\biguplus_j B_j$.

פונקציית התועלת: $P^S_y(B)=\frac{1}{n_S(B)}\sum_{x_i\in B}1_{\{y_i=y\}}$, לפי עיקרון ה-ERM מי שממקסם את ההסתברות זה כלל הרוב הפשוט.

גידול עץ: תועלת החיתוך-
$$g_i(t)=P^S_{\hat y(S(B_{>t}))}(B_{>t})+P^S_{\hat y(S(B_{<t}))}(B_{<t})$$
$$\arg\max_{i,t}g_i(t)$$
זמן ריצה: $O(md\cdot 2^k)\overset{2^k\le m}{=}O(m^2d)$.

PAC & Agnostic PAC

Generalization Error: $L_{\mathcal{D},f}(h)=\Pr_{x\sim\mathcal{D}}[h(x)\neq f(x)]$

Probably Approximately: אלגוריתם למידה $\mathcal{A}$ בעל חסם תחתון $\varepsilon>0$ (דיוק) על ההפסד בהסתברות קטנה מ-$\delta>0$ (ביטחון).

PAC Learnable: נאמר שמחלקת היפותזות $\mathcal{H}$ היא PAC Learnable אם קיימת פונקציה $\widetilde m_{\mathcal{H}}:(0,1)^2\to\mathbb{N}$ ואלגוריתם למידה $\mathcal{A}_m:(\mathcal{X}\times\mathcal{Y})^m\to\mathcal{H}$ **בעל התכונה:**
● לכל $\varepsilon,\delta\in(0,1)$ ● לכל התפלגות $\mathcal{D}$ מעל $\mathcal{X}$ ● לכל פונקציה תיוג $f:\mathcal{X}\to\mathcal{Y}$ (המקיימת את הנחת הריאליזביליות)
כאשר נריץ את $\mathcal{A}$ על $m\ge\widetilde m_{\mathcal{H}}(\varepsilon,\delta)$ דגימות $i.i.d$ מעל $\mathcal{D}$ מתוייגות ע"י f, מתקיים:
$$\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}\left[L_{\mathcal{D},f}(\mathcal{A}(S))>\varepsilon\right]<\delta$$
עבור loss חסום, התכונה שקולה: $\lim_{m\to\infty}\mathbb{E}_{S\overset{iid}{\sim}\mathcal{D}^m}[L_{\mathcal{D},f}(\mathcal{A}(S))]=0$

Sample Complexity: $\widetilde m_{\mathcal{H}}(\varepsilon,\delta)$ המינימלי.

אין ארוחות חינם: יהי $\mathcal{X}$ מרחב דגימות אינסופי, ונקבע $\varepsilon<\frac12$. קיימת $\delta>0$ כך שלכל אלגוריתם למידה $\mathcal{A}$ וקורפוס אימון בגודל m, קיימת התפלגות $\mathcal{D}(\mathcal{X})$ מעל $\mathcal{X}$ ופונקצייה $f:\mathcal{X}\to\mathcal{Y}$ כך ש:
$$\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}\left[L_{\mathcal{D},f}(\mathcal{A}(S))>\varepsilon\right]\ge\delta$$
Realizability Assumption: $\exists h\in\mathcal{H}\ \Pr_{x\sim\mathcal{D}}[h(x)=f(x)]=1$

VC-Dimension: תהי $\mathcal{H}\subset\{\pm1\}^{\mathcal{X}}$ מחלקת היפותזות. נסמן $\mathcal{H}_C=\{h\restriction_C:h\in\mathcal{H}\}$:
$$VCdim(\mathcal{H})=\max\{|C|:C\subset\mathcal{X}\ and\ |\mathcal{H}_C|=2^{|C|}\}$$

הכללת המושגים:
● התפלגות $\mathcal{D}$ עבור $\mathcal{X}\times\mathcal{Y}$ ● Approximately Correct עבור $h\in\mathcal{H}$ המקיימת $L_{\mathcal{D}}(h)\le\min_{h'\in\mathcal{H}}L_{\mathcal{D}}(h')+\varepsilon$ ● הכללת השגיאה ע"י $L_{\mathcal{D}}(h)=\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x),y)]$

Agnostic-PAC Learnable: נאמר שמחלקת היפותזות $\mathcal{H}$ היא Agnostic-PAC Learnable ביחס ל-$\ell:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to[0,\infty)$ אם קיימת פונקציה $\widetilde m_{\mathcal{H}}:(0,1)^2\to\mathbb{N}$ ואלגוריתם למידה $\mathcal{A}_m:(\mathcal{X}\times\mathcal{Y})^m\to\mathcal{H}$ בעל התכונה:
● לכל $\varepsilon,\delta\in(0,1)$ ● לכל התפלגות $\mathcal{D}$ מעל $\mathcal{X}\times\mathcal{Y}$
כאשר נריץ את $\mathcal{A}$ על $m\ge\widetilde m_{\mathcal{H}}(\varepsilon,\delta)$ דגימות $i.i.d$ מעל $\mathcal{D}$, מתקיים:
$$\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}\left[L_{\mathcal{D},f}(h_S)\le\min_{h\in\mathcal{H}}L_{\mathcal{D}}(h)+\varepsilon\right]\ge1-\delta$$

המשפט היסודי של הלמידה הסטטיסטית:

תהי $\mathcal{H}\subset\{\pm1\}^{\mathcal{X}}$ מחלקת היפותזות ונסמן $d=VCdim(\mathcal{H})$. אז $\mathcal{H}$ למידה PAC אמ"ם $\mathcal{H}$ למידה Agnostic-PAC אמ"ם $d<\infty$. במקרה זה:
● קיימים קבועים אוניברסליים C_1,C_2 כך ש-$\mathcal{H}$ למידה PAC עם סיבוכיות דגימות:
$$C_1\frac{d+\log\left(\frac1\delta\right)}{\varepsilon}\le m_{\mathcal{H}}(\varepsilon,\delta)\le C_2\frac{d\log\left(\frac1\varepsilon\right)+\log\left(\frac1\delta\right)}{\varepsilon}$$
● קיימים קבועים אוניברסליים C_1,C_2 כך ש-$\mathcal{H}$ למידה Agnostic-PAC עם סיבוכיות דגימות:
$$C_1\frac{d+\log\left(\frac1\delta\right)}{\varepsilon^2}\le m_{\mathcal{H}}(\varepsilon,\delta)\le C_2\frac{d+\log\left(\frac1\delta\right)}{\varepsilon^2}$$
● את החסם העליון ניתן להשיג ע"י לומד ERM.

ε-מייצגת: עבור $\ell,\mathcal{H},\mathcal{D}$, קבוצת דגימות S המקיימת:
$$\forall h\in\mathcal{H}\ |L_S(h)-L_{\mathcal{D}}(h)|<\varepsilon$$
למה 1: תהי S קבוצת דגימות ε-מייצגת עבור $\ell,\mathcal{H},\mathcal{D}$. תהי $h_S\in ERM_{\mathcal{H}}(S)$ אז $L_{\mathcal{D}}(h_S)\le\min_{h\in\mathcal{H}}L_{\mathcal{D}}(h)+\varepsilon$.

התכנסות אחידה: למחלקת היפותזות $\mathcal{H}$ יש תכונת התכנסות אחידה אם קיימת $m^{UC}_{\mathcal{H}}:(0,1)^2\to\mathbb{N}$ כך שלכל $\varepsilon,\delta\in(0,1)$ ולכל התפלגות $\mathcal{D}$ מעל $\mathcal{X}\times\mathcal{Y}$ מתקיים:
$$\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}[S\ is\ \varepsilon-representative]\ge1-\delta$$
טענה: אם $\mathcal{H}$ בעלת תכונת התכנסות אחידה עם פונקציה $m^{UC}_{\mathcal{H}}:(0,1)^2\to\mathbb{N}$ אז $\mathcal{H}$ למידה Agnostic-PAC עם סיבוכיות דגימות $m_{\mathcal{H}}(\varepsilon,\delta)\le m^{UC}_{\mathcal{H}}\left(\frac\varepsilon2,\delta\right)$.

הוכחת המשפט (המקרה הכללי)

1. אם $VCdin(\mathcal{H})<\infty$ אז $|\mathcal{H}_C|$ גדל פולינומיאלית ב-$|C|$.

גדילה פולינומיאלית: תהי $\mathcal{H}\subset\mathcal{Y}^{\mathcal{X}}$. נגיד ש-$\mathcal{H}_C$ גדלה פולינומיאלית ב-$|C|$ אם קיימים $m_0\in\mathbb{N}$, $b,\beta>0$ עבורם:
$$\forall m>m_0\ \tau_{\mathcal{H}}(m):=\max_{\substack{C\subset\mathcal{X}\\|C|=m}}|\mathcal{H}_C|\le b\cdot m^\beta$$
למה 3: אם $m>VCdim(\mathcal{H})$ אז $\tau_{\mathcal{H}}(m)\le\left(\frac{em}{d}\right)^d$.

2. אם $|\mathcal{H}_C|$ גדל פולינומיאלית ב-$|C|$, אז $\mathcal{H}$ מקיים את תכונת ההתכנסות האחידה.

נגדיר: $F^{\mathcal{D}}_m(S):=\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h)-L_S(h)|$

צ"ל: $\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}[F^{\mathcal{D}}_m(S)>\varepsilon]<\delta$. לפי מרקוב מספיק להראות:
$$\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}[F^{\mathcal{D}}_m(S)>\varepsilon]\le\frac{\mathbb{E}_{S\overset{iid}{\sim}\mathcal{D}^m}[F^{\mathcal{D}}_m(S)]}{\varepsilon}\le\frac{\alpha_m}{\varepsilon}$$
למה 2:
$$\mathbb{E}_{S\overset{iid}{\sim}\mathcal{D}^m}[F^{\mathcal{D}}_m(S)]\le O\left(\frac{\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}\right)+o(m)$$

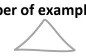
הוכחת המשפט (המקרה הסופי)

$$\begin{aligned}\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}[F^{\mathcal{D}}_m(S)>\varepsilon]&\overset{def}{=}\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}[\exists h\in\mathcal{H}\ |L_{\mathcal{D}}(h)-L_S(h)|>\varepsilon]\\&\overset{union\ bound}{\le}\sum_{h\in\mathcal{H}}\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}[|L_{\mathcal{D}}(h)-L_S(h)|>\varepsilon]\\&\le|\mathcal{H}|\max_{h\in\mathcal{H}}\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}[|L_{\mathcal{D}}(h)-L_S(h)|>\varepsilon]\\&\overset{Hoeffding}{\le}|\mathcal{H}|\cdot2e^{-2m\varepsilon^2}\end{aligned}$$
אם ניקח $m\ge\frac{\log\left(2|\mathcal{H}|/\delta\right)}{2\varepsilon^2}$ נקבל את הנדרש.

מחלקות סופיות

חוק עקבי: חוק $h\in\mathcal{H}$ המקיים $y_i=h(x_i)$ עבור $i\in[m]$.

משפט: כל מחלקת היפותזות $\mathcal{H}$ מגודל סופי היא למידה PAC ע"י שימוש בעקרון ה-ERM ובעלת סיבוכיות דגימה:
$$m_{\mathcal{H}}(\epsilon,\delta)\le\frac{\log\left(|\mathcal{H}|/\delta\right)}{\epsilon}$$
Number of examples (m)



Accuracy (ε) $\frac{\log\left(|\mathcal{H}|/\delta\right)}{m}$ **Confidence (δ)** $\frac{|\mathcal{H}|}{e^{m\cdot\epsilon}}$

טענה: עבור $\mathcal{H}$ סופית, $VCdim(\mathcal{H})\le\log_2|\mathcal{H}|$.

Bagging & Boosting

טענה: בהינתן מ"מ $X_1,\dots,X_T$ בעלי שונות σ^2 וקורלציה $corr(X_i,X_j)=\rho$. השונות של $\bar X=\frac1T\sum_{t=1}^TX_t$ היא:
$$\rho\cdot\sigma^2+(1-\rho)\cdot\frac{\sigma^2}{T}$$
Meta-Algorithms: בחירה ע"י רוב:
● קלסיפיקציה $h(x)=sign(\sum_t h_t(x))$
● רגרסיה $h(x)=\frac1T\sum_t h_t(x)$

חסרונות: ● צריך לאמן T מודלים. ● בשביל חיזוי צריך לשמור T מודלים. ● יותר קשה לפרש מדוע האלגוריתם חזה משהו.

לא להשתמש כאשר הצלחת הלומד קטנה מחצי!

The Bootstrap

בהינתן קבוצת דגימות S אנו יוצרים B קבוצות חדשות ע"י דגימה חוזרת מתוך S בגודל m פעמים:
$$S^b=\{(x_i^b,y_i^b)\}_{i=1}^m$$
Empirical Distribution: התפלגות $\hat{\mathcal{D}}_S$ המושרת ע"י S מעל $\mathcal{X}\times\mathcal{Y}$ על תת-קבוצה $C\subset\mathcal{X}\times\mathcal{Y}$:
$$\hat{\mathcal{D}}_S((X,Y)=(x,y))=\begin{cases}\frac1m&(x,y)\in S\\0&otherwise\end{cases},\ \hat{\mathcal{D}}_S(C)=\frac{|C\cap S|}{m}$$
טענה: מ"מ $X_1,\dots,X_T$ ב"ת עם התפלגות F כלשהי מעל $\mathbb{R}$. יהי x_i ערך דגימה של X_i, אז ה-CDF (הצטברות משותפת) היא פונקציית מדרגות העולה מ-0 ל-1 בקפיצות של $\frac1m$ בכל x_i.

טענה: עבור m גדול דיו, בערך 37% מהנקודות נותרות מחוץ לתהליך ה-bootstrap שנוצר מ-S.

Bagging (הורדת ה-Variance, ה-Bias נותר זהה)
עבור אלגוריתם למידה $\mathcal{A}$ בסיסי בוחרים T ויוצרים מדגמים $S^1,\dots,S^T$ כל אחד בגודל m. מאמנים T היפותזות $h_{S^1},\dots,h_{S^T}$.
חיזוי דגימה חדשה:
$$h_{bag}(x)=sign\left(\sum_{t=1}^T h_{S^t}(x)\right)$$
De-correlation: מכיוון שניתן להוריד את השונות רק עד מידה מסויימת, אנו מבצעים תהליך דה-קורלציה. למשל בעצים נבחר פרמטר k ובכל עת נבחר k פרמטרים שרק מתוכם ניתן לפצל.

Boosting (הורדת ה-Bias, ה-Variance עולה לאט)
ע"י עדכון מרחב המדגם S^t (הוספת משקולות) כך שדגימות בהן טעינו יקבלו משקל גדול יותר (יוצר התפלגות D^t). [או שניתן לדגום מחדש מתוך התפלגות D^t, תמיד זמן אך יכול ליצור כפילויות דגימה]

Adaboost Algorithm:

Initialize: $D^1=\left(\frac1m,\dots,\frac1m\right)$

Loop: $h_t=\mathcal{A}(D^t,S)$

Update:
● $\sum_{j=1}^mD_j^{t+1}\cdot1_{[h_t(x_j)\neq y_j]}=\frac12\Rightarrow w_t=\frac12\log\left(\frac1{\varepsilon_t}-1\right)$
($\varepsilon_t=\sum_{j=1}^mD_j^t\cdot1_{[h_t(x_j)\neq y_j]}$,סיכון אמפירי ממושקל)
● $D_i^{t+1}=\frac{D_i^t\cdot e^{-w_ty_ih_t(x_i)}}{\sum_{j=1}^mD_j^t\cdot e^{-w_ty_jh_t(x_j)}}$

חיזוי דגימה חדשה:
$$h_{boost}(x)=sign\left(\sum_{t=1}^Tw_th_t(x)\right)$$
γ-weak-learner: אלגוריתם למידה $\mathcal{A}$ הוא לומד-γ חלש עבור מחלקת היפותזות $\mathcal{H}$ אם קיימת פונקציה $m_{\mathcal{H}}:(0,1)\to\mathbb{N}$ כך שלכל $\delta\in(0,1)$, לכל התפלגות $\mathcal{D}$ מעל מרחב מדגם $\mathcal{X}$, ולכל פונקציית תיוג $f:\mathcal{X}\to\{\pm\}$, אם הנחת הריאליזביליות מתקיימת (ביחס ל-$\mathcal{H},\mathcal{D},f$), אז כאשר נריץ את $\mathcal{A}$ על $m\ge m_{\mathcal{H}}(\delta)$ דגימות iid מתוך $\mathcal{D}\times f$ האלגוריתם יחזיר היפותזה $h_S=\mathcal{A}(S)$ כך ש:
$$\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}\left[L_{\mathcal{D},f}(h_S)\le\frac12-\gamma\right]\ge1-\delta$$
γ-weak-learnable: מחלקה $\mathcal{H}$ היא למידה-γ חלשה אם קיים $\mathcal{A}$ כנ"ל.

משפט: למידות חלשה $\Leftarrow$ $VCdim(\mathcal{H})<\infty$.

משפט: תהי S קבוצת דגימות. נניח שבכל איטרציה של Adaboost הלומד הבסיסי מחזיר כלל חיזוי (היפותזה h_t) עבודה הסיכון האמפירי הממושקל מקיים
$$\varepsilon_t=\sum_{j=1}^mD_j^t\cdot1_{[h_t(x_j)\neq y_j]}\le\frac12-\gamma$$
אז כלל הסיכון האמפירי (הלא ממושקל) של כלל ההחלטה של Adaboost מקיים:
$$L_S(h_{boost})\equiv\frac1m\sum_{j=1}^m1_{[h_{boost}(x_j)\neq y_j]}\le e^{-2\gamma^2T}$$

	Bagging	Boosting
מקביליות	במקביל	טורי
מבנה הנתונים	Bootstrap training samples	Bootstrap ממושקל או קבוצה S המקורית עם כלל ERM ממושקל
De-correlation	מומלץ	אין צורך
Overfitting	לא קורה	סיכון
איזה מודל להשתמש	עצים עמוקים	עצים רדודים
השפעה	מפחית Variance	מפחית Bias

Regression & Regularization

עבור $\lambda\ge0$ נגדיר לומד $\mathcal{A}_\lambda:S\mapsto h_S$ ע"י:
$$h_S=\arg\min_{h\in\mathcal{H}}\mathcal{F}_S(h)+\lambda\cdot\mathcal{R}(h)$$
Regularization Term: $\mathcal{R}$ מהנוסחה לעיל הוא תנאי הרגולריזציה.
● עבור $\lambda=0$ אין התייחסות כלל למורכבות ההיפותזה שנבחרה. ה-Bias נמוך אך ה-Variance גבוה.
● עבור $\lambda\to\infty$ נעדיף היפותזה כמה שיותר פשוטה ללא קשר לפונקציית המטרה.

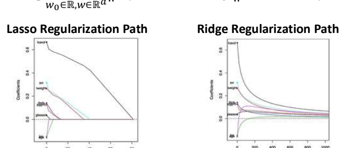
Regression Trees

עבור עץ רגרסיה $h(x)=\sum_{j=1}^Nc_j\cdot1_{[x\in B_j]}$ נגדיר את $|T|$ להיות מספר העלים, ואת T_0 להיות העץ המתקבל מתהליך הגידול:
$$\min_{T\subseteq T_0}L_S(T)+\lambda\cdot|T|$$

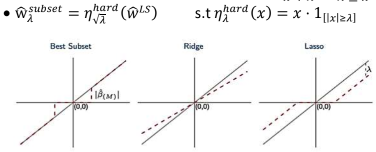
Modern Regression

● **Ridge Regression**
$$\arg\min_{w_0\in\mathbb{R},w\in\mathbb{R}^d}\left\|w_0\cdot\vec1+X^Tw-y\right\|^2+\lambda\cdot\|w\|_2^2$$
יש לפתור את המערכת $Xy=(XX^T+\lambda I)w$. פתרון יהיה $\hat w_\lambda=U\Sigma^\lambda V^Ty$ כאשר $\Sigma^\lambda=diag\left(\frac{\sigma_i}{\sigma_i^2+\lambda}\right)$.

● **Lasso** – בעל תכונת דלילות.
$$\arg\min_{w_0\in\mathbb{R},w\in\mathbb{R}^d}\left\|w_0\cdot\vec1+X^Tw-y\right\|^2+\lambda\cdot\|w\|_1$$
● **Best Subset**
$$\arg\min_{w_0\in\mathbb{R},w\in\mathbb{R}^d}\left\|w_0\cdot\vec1+X^Tw-y\right\|^2+\lambda\cdot\|w\|_0$$



טענה: עבור $w\in\mathbb{R}^d$ ומטריצה $XX^T=I$ (תכנון אורתוגונלי):
● $\hat w_\lambda^{ridge}=\frac1{1+\lambda}\hat w^{LS}$
● $\hat w_\lambda^{lasso}=\eta_\lambda^{soft}(\hat w^{LS})$ s.t $\eta_\lambda^{soft}(x)=\begin{cases}x-\lambda & x\ge\lambda\\0&-\lambda<x<\lambda\\x+\lambda&-\lambda\ge x\end{cases}$
● $\hat w_\lambda^{subset}=\eta^{hard}_{\sqrt\lambda}(\hat w^{LS})$ s.t $\eta_\lambda^{hard}(x)=x\cdot1_{[|x|\ge\lambda]}$



Regularized Logistic Regression

$$\arg\min_{w_0,w}\left[\sum_{i=1}^m\left[\log\left(1+e^{w_0+\langle x_i,w\rangle}\right)-y_i(w_0+\langle x_i,w\rangle)\right]+\lambda\cdot\|w\|_1\right]$$

Model Selection

k-fold Cross Validation: עבור $i=1,\dots,k$ נאמן את המודל ה-i במספר על כל הדאטה מלבד החלק ה-i. נחשב את ה-loss של המודל ה-i ע"פ החלק ה-i. נחזיר את הממוצע וסטיית התקן של התוצאות.
$$\sigma=\sqrt{\frac1{n-1}\sum_{i=1}^n(x_i-\bar x)^2}$$

Bootstrap: נדגום B קבוצות S^b, ונאמן את האלגוריתם עליהם. בכדי לבחון את הביצועים נשתמש ב-$T^b = S \setminus S^b$ ונמצע כמקודם.

טעויות נפוצות:

1. **Over-estimating generalization error** – כאשר מאמנים על $\frac{k-1}{k}\cdot m$ כמות לא מספקת של דגימות, אז לפי PAC נעריך בחסר את השגיאה.
2. **Under-estimating generalization error** – כאשר אנו מעריכים מודל ע"פ נתונים עליהם אומן (מצב של Overfitting), או כאשר על הדאטה בוצע עיסוי המתאים למודל.

<u>השגיאה בעת שימוש בולידציה</u>

$$L_{\mathcal{D}}(h_S) = \underbrace{L_{\mathcal{D}}(h_S) - L_{\mathcal{V}}(h_S)}_{A} + \underbrace{L_{\mathcal{V}}(h_S) - L_S(h_S)}_{B} + \underbrace{L_S(h_S)}_{C}$$

A. שגיאת ההכללה. ניתנת לחסימה תחת ההנחה שפונקציית ה-loss חסומה. **B.** עבור B גדול ו-C קטן ככל הנראה Overfitting. **C.** עבור C גדול ככל הנראה Underfitting.

טענה: לכל $h \in \mathcal{H}$ ולכל $\delta \in (0,1)$:

$$\Pr\left[|L_{\mathcal{V}}(h) - L_{\mathcal{D}}(h)| \le \sqrt{\frac{\ln(2/\delta)}{m_{\mathcal{V}}}}\right] \ge 1-\delta$$

עבור $\mathcal{H}_1 \subseteq \cdots \subseteq \mathcal{H}_k$ סופיים.

Standard Method: עבור $h^* \in ERM_{\mathcal{H}_k}(S_{all})$:

$$\Pr\left[L_{\mathcal{D}}(h^*) \le \min_{h\in\mathcal{H}_k} L_{\mathcal{D}}(h) + \sqrt{\frac{21\left(2|\mathcal{H}_k|/\delta\right)}{m}}\right] \ge 1-\delta$$

Model Selection: עבור $h^* \in \arg\min_{h\in\mathcal{H}_k} L_{\mathcal{D}}(h) \subseteq \mathcal{H}_j \subseteq \cdots \subseteq \mathcal{H}_k$

$$\Pr\left[L_{\mathcal{D}}(h^*) \le \min_{h\in\mathcal{H}_k} L_{\mathcal{D}}(h) + \sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m}\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}\right] \ge 1-\delta$$

<u>בחירת פיטצ'רים</u>

1. **מתאם פירסון**

$$R = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})\sum_i (y_i-\bar{y})}}$$

2. **Forward-Stepwise Selection** - אלגוריתם חמדן שמתחיל מהחותך (intercept) ומוסיף פיטצ'רים עד לכמות k הנקבעת.
3. **Backward-Stepwise Selection** - אלגוריתם המתחיל מכל הפרמטרים ומוריד כל פעם פרמטר עם המתאם הקטן ביותר.
4. **Regularization** - שיטה כללית בה מגבילים את כמות הפרמטרים.

בנוסף ניתן לבצע **שינוי של המידע**, למשל מרכוז שלו סביב הראשית, שינוי הטווח (Scaling), ביצוע Standardization ועוד...

Unsupervised Learning

<u>דוגמאות לשימוש:</u> **1.** חשיפה של מבנה במימד נמוך. **2.** קלסיפיקציה. **3.** זיהוי אנומליות.

<u>PCA</u>

$$\arg\min_{U\in\mathbb{R}^{d\times k}, W\in\mathbb{R}^{k\times d}} \sum_{i=1}^{m} \|x_i - UWx_i\|^2$$

משפט: תהי $A = \sum_{i=1}^m x_i x_i^T$, ויהיו $u_1,\dots,u_n$ הו"ע של A (מסודרים בסדר יורד), אז $U = W^T = [u_1 \cdots u_k]$ פתרון לבעיית ה-PCA.

- $P = \sum_{i=1}^k u_i u_i^T$ נקראת מטריצת הטלה, היא סימטרית והיא קירוב טוב ביותר ל-x בתת-מרחב:
$$\|x-v\|_2 \ge \|x - Px\|_2 \quad \forall v \in span\{u_1,\dots,u_k\}$$
- היא שונה מ-LS בו מטילים את הערך y בלבד.

ע"י הזזה של הנקודות $\tilde{x}_i = x_i - \bar{x}$ אנו מקבלים ב-A את מטריצה השונות: $A = \frac{1}{m-1}\sum_i (x_i-\bar{x})(x_i-\bar{x})^T$.

הגדרות: תהי מטריצת השונות של דגימות אימון $x_1,\dots,x_m$ כמוגדר לעיל, ויהיו $u_1,\dots,u_d$ הו"ע של מטריצת השונות המתאימים לע"ע $\lambda_1 \ge \cdots \ge \lambda_d \ge 0$.

- המספר λ_i נקרא <u>הערך המוביל</u> ה-i של $x_1,\dots,x_m$.
- הוקטור u_i נקרא <u>הוקטור המוביל</u> ה-i של $x_1,\dots,x_m$.

הורדת מימד תהיה $U^T x_i$. לאחר מכן כל וקטור במרחב ניתן לייצוג ע"י הקירוב $\bar{x} + \sum \alpha_i u_i$.

```
PCA Algorithm:
Input: X ∈ ℝ^{m×d}
Eval:
  • if m > d:
      o A = X^T X
      o u_1, …, u_k eigenvectors of A
  • else (m ≤ d):
      o B = X X^T
      o v_1, …, v_k eigenvectors of B
      o denote u_i = (1/‖X^T v_i‖) X^T v_i
  • return u_1, …, u_k
```

בשביל לבחור את k ניתן לייצר Scree Plot (גרף בו מציגים את הע"ע בסדר יורד).

<u>Clustering</u>

נגדיר חלוקה של המידע $x_1,\dots,x_m$ ל-k מחלקות $\biguplus_j C_j$. נגדיר את פונקציית המחיר (ע"פ פונקציית מרחק d) ע"י:

$$G(C_1,\dots,C_k) = \min_{\mu_1,\dots,\mu_k} \sum_{j=1}^{k} \sum_{x\in C_j} d(x,\mu_j)^2$$

Centroid: (של C_j) הערך $\arg\min_{\mu_j} \sum_{x\in C_j} d(x,\mu_j)^2$.

טענה: אם $\mathcal{X} = \mathbb{R}^d$ ו-d היא הנורמה האוקלידית, אז ה-*centroid* הוא הממוצע $\mu_j = \frac{1}{|C_j|}\sum_{x\in C_j} x$.

```
k-means Algorithm:
Input: x_1, … x_m and k ∈ ℕ
Step 0: choose initial μ_1, …, μ_k
Until convergence:
  • set C_j to be the points x_i closer to
    μ_j than to any other centroid.
  • update μ_j to centroid of C_j:
        μ_j = (1/|C_j|) Σ_{x∈C_j} x
```

החלוקה המושרית מ-$\mu_1,\dots,\mu_k$ נקראת Voronoi cells.

תכונות: השונות בכל C_j יורדת בכל איטרציה. האלגוריתם תמיד מתכנס. האלגוריתם מתכנס למינימום מקומי (בהתאם לנקודות ההתחלה).

Convex Optimization

פונקציה קמורה: תהי C קבוצה קמורה. פונקציה $f: C \to \mathbb{R}$ נקראת קמורה אם:

$$\forall u,v \in C, \alpha\in[0,1] \;\; f(\alpha v + (1-\alpha)u) \le \alpha f(u) + (1-\alpha) f(v)$$

משפט: חיתוך של קבוצות קמורות הוא קמור. $\bigcap_\alpha A_\alpha$.

אפיגרף: $epi(f) = \{(x,t) | f(x) \le t\}$

משפט: פונקציה f קמורה **אם"ם** $epi(f)$ קבוצה קמורה **אם"ם** לכל $x,y \in dom(f)$ מתקיים $f(y) \ge f(x) + \nabla f(x)^T (y-x)$ (תנאי מסדר ראשון) **אם"ם** $\nabla^2 f(x) \succcurlyeq 0$ (תנאי מסדר שני).

תכונות השומרות על קמירות:

1. **סכום חיובי** - עבור $f_1,\dots,f_n$ קמורות, $\alpha_1,\dots,\alpha_n \ge 0$, אז $f = \sum_i \alpha_i f_i$ קמורה.
2. **הרכבה עם אפינית** - $g(x) = f(Ax+b)$ קמורה על $\{x | Ax+b \in dom(f)\}$.
3. **מקסימום** - אם לכל $y \in A$ מתקיים $f(x,y)$ קמורה ב-x, אז $g(x) = \sup_{y\in A} f(x,y)$ קמורה.
4. **מינימום** - אם $f(x,y)$ קמורה במשותף ($\nabla^2 f \succcurlyeq 0$) ו-$S$ קמורה, אז $g(x) = \inf_{y\in S} f(x,y)$ קמורה. (לא בקורס).

<u>תכונות - מינימום</u>

משפט: אם f קמורה, כל מינימום מקומי הוא גלובלי.
משפט: תהי f קמורה ודפרנציבילית ו-$w \in dom(f)$, אז:
$$\forall u \;\; f(u) \ge f(w) + \langle \nabla f(w), u-w\rangle$$

<u>תכונות – סאב-גרדיאנט</u>

סאב-גרדיאנט: וקטור v הוא סאב-גרדיאנט של w אם:
$$\forall u \;\; f(u) \ge f(w) + \langle v, u-w\rangle$$
סימון: $\partial f(w)$ זוהי קבוצת כל סאב-הגרדיאנטים של f בנקודה w.

למה: f קמורה אם"ם לכל נקודה $w \in dom(f)$ מתקיים $\partial f(w) \ne \emptyset$. אם f דפרנציבילית ב-x אז $\partial f(x) = \{\nabla f(x)\}$.

טענה: נניח $f_i: V\to\mathbb{R}$ קמורות, ותהי $f(v) = \max_i f_i(v)$. כמו כן עבור $u \in V$ נסמן $j \in \arg\max_i f_i(u)$, אז $\partial f_j(u) \subseteq \partial f(u)$.

מסקנה: נניח $f: V\to\mathbb{R}$ קמורה. תהי $w \in V$ עבורה $\vec{0} \in \partial f(u)$, אז w מינימום גלובלי של f.

<u>תכונות - ליפשיץ</u>

ליפשיץ: פונקציה $f: C\to\mathbb{R}$ נקראת ρ-ליפשיץ אם:
$$\forall w_1,w_2 \in C \;\; |f(w_1) - f(w_2)| \le \rho\|w_1 - w_2\|$$
למה: אם f קמורה, אז f היא ρ-לפישיץ אם"ם הנורמה של כל סאב-גרדיאנט של f הוא לכל היותר ρ.

<u>אופטימיזציה קמורה</u>

בעיית אופטימיזציה:
$$\min_x \; f_0(x) \quad s.t \;\; f_i(x) \le b_i \;\; \forall i = 1,\dots,n$$

בעיית אופטימיזציה קמורה: בעיית אופטימיזציה עם f_i קמורות.
בעיית תכנון לינארי: בעיית אופטימיזציה עם f_i לינאריות.

בעיית למידה קמורה: בעיית למידה $\mathcal{H},\ell$ מעל $\mathcal{X}\times\mathcal{Y}$ תקרא קמורה אם מחלקת ההיפותזות $\mathcal{H}$ היא קבוצה קמורה, ולכל $(x,y) \in \mathcal{X}\times\mathcal{Y}$ הפונקציה $\ell(h(x),y)$ קמורה ב-h.
דוגמה: $ERM_{\mathcal{H}}$.

טענה: לא כל הבעיות הקמורות מעל $\mathbb{R}^d$ הן למידות-PAC.
טענה: בעיה קמורה מעל $\mathbb{R}^d$, כך ש-$\mathcal{H}$ חסומה ו-ℓ ליפשיץ, היא למידה-PAC.

- **$\mathcal{H}$ חסומה** – קיים B עבורו $\forall w \in \mathcal{H} \;\; \|w\| \le B$.
- **ℓ ליפשיץ** – הפונקציה $\ell(h(x),y)$ קמורה ב-h והיא ρ-ליפשיץ.

במקרה זה סיבוכיות הדגימה תלויה רק ב-ϵ,δ,B,ρ.

```
Sub-gradient Descent Algorithm:
initialize: w^(1) = 0
for t = 1, …, T:
  • Choose v_t ∈ ∂f(w^(t))
  • w^(t+1) ← w^(t) − η · v_t
return w̄ = (1/T) Σ_{t=1}^T w^(t)
```

משפט: תהי f פונקציה קמורה, ρ-ליפשיץ, ויהי $w^* \in \arg\min_w f(w)$. אם נריץ את האלגוריתם Sub-GD על f במשך T צעדים עם $\eta = \frac{\|w^*\|}{\rho\sqrt{T}}$, אז הוקטור $\bar{w}$ המוחזר מקיים:
$$f(\bar{w}) \le f(w^*) + \frac{\|w^*\|\rho}{\sqrt{T}}$$

מסקנה: תהי f פונקציה קמורה, ρ-ליפשיץ, ויהי $w^* \in \arg\min_w f(w)$. לכל $\varepsilon > 0$ אם נריץ את האלגוריתם Sub-GD על f במשך $\frac{\|w^*\|^2\rho^2}{\varepsilon^2}$ צעדים עם $\eta = \frac{\|w^*\|}{\rho\sqrt{T}}$, אז הוקטור $\bar{w}$ המוחזר מקיים:
$$f(\bar{w}) \le f(w^*) + \varepsilon$$

מסקנה: האלגוריתם Sub-GD צריך $\frac{\|w^*\|^2\rho^2}{\varepsilon^2}$ איטרציות בכדי להתכנס.

```
Stochastic Gradient Descent Algorithm:
initialize: w^(1) = 0
for t = 1, …, T:
  • Choose (x, y) ~ 𝒟
  • Choose v_t ∈ ∂ℓ(w^(t), (x, y))
  • w^(t+1) ← w^(t) − η · v_t
return w̄ = (1/T) Σ_{t=1}^T w^(t)
```

למה: תהי בעיית למידה קמורה-ליפשיץ-חסומה עם פרמטרים ρ, B. אז לכל $\epsilon > 0$, אם נריץ את האלגוריתם SGD למזעור $L_{\mathcal{D}}(w)$ עם מספר איטרציות $T \ge \frac{B^2\rho^2}{\varepsilon^2}$ ועם $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, אז הפלט של SGD מקיים:
$$\mathbb{E}[L_{\mathcal{D}}(\bar{w})] \le \min_{w\in\mathcal{H}} L_{\mathcal{D}}(w) + \varepsilon$$
טענה: אלגוריתם $\mathcal{A}$ הלומד בעזרת SGD הוא לומד-חלש.

כדאי לזכור

טבלת VCdim:

Class	VCdim
Axis aligned rectangles	4
Homogeneous Halfspaces	d
Non- Homogeneous Halfspaces	$d+1$
Polynomial Thresholds (Pol_r)	$\le \binom{d+r-1}{d}$
k-interval	$2k$
Conjuctions	d
sin	∞

תזכורת: בשביל $VCdim(\mathcal{H}) = d$ צריך להראות שקיימת מחלקה מגודל d מנותצת, ושכל מחלקה בגודל $d+1$ לא ניתן לנתץ.

ההבדלים בשגיאות:

- **Generalization error:** $L_{\mathcal{D}}(h)$
- **Training error / empirical error:** $L_S(h)$
- **Approximation error:** $\min_h L_{\mathcal{D}}(h)$
- **Estimated error:** $L_{\mathcal{D}}(h) - \min_{h'} L_{\mathcal{D}}(h')$

<u>לשים לב:</u> כאשר נגדיל את T ב-Adaboost (כלומר את מספר הלומדים החלשים) אנו נשפר את ה-Approximation error (ובהתאם את ה-Estimated error).

Bias-Variance Tradeoff: יהי $h^* \in \arg\min_h L_{\mathcal{D}}(h)$, $h_S = \mathcal{A}(S)$. אזי:

$$L_{\mathcal{D}}(h_S) = \underbrace{L_{\mathcal{D}}(h^*)}_{\varepsilon_{approximation}} + \underbrace{L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*)}_{\varepsilon_{estimation}}$$

KNN		
Low Bias	Small $\xrightarrow{k}$ Big	High Bias
Low Var	Big $\xleftarrow{k}$ Small	High Var

Decision Tree		
Low Bias	Deep $\xleftarrow{height}$ Shallow	High Bias
Low Var	Shallow $\xrightarrow{height}$ Deep	High Var

Linear Regression		
Low Bias	$(d \sim m)$ $\quad (d \ll m)$	High Bias
Low Var	$(d \ll m)$ $\quad (d \sim m)$	High Var

Regularization		
Low Bias	$0 \xrightarrow{\lambda} \infty$	High Bias
Low Var	$\infty \xleftarrow{\lambda} 0$	High Var

Bagging		
Bias	Dosen't change	Bias
Low Var	Big(∞) $\xleftarrow{T}$ Small	High Var

Adaboost		
Low Bias	Big(∞) $\xleftarrow{T}$ Small	High Bias
Low Var	Small $\xrightarrow{T}$ Big(∞)	Var goes up a bit

Clustering		
Low Bias	Big $\xleftarrow{k}$ Small	High Bias
Low Var	Small $\xrightarrow{k}$ Big	High Var

<u>אלגברה לינארית</u>

טענה: עבור מטריצה $A \in \mathbb{R}^{m\times n}$:
$$\mathbb{R}^m = \mathcal{C}(A) \oplus \mathcal{N}(A^T), \qquad \mathbb{R}^n = \mathcal{C}(A^T) \oplus \mathcal{N}(A)$$

משפט: עבור מטריצה $A \in \mathbb{R}^{m\times n}$ הבאים שקולים:

1. A מדרגה מלאה ($rank(A) = m$)
2. $\det(A) \ne 0$
3. $im(A) = \mathbb{R}^m$
4. $\ker(A) = \{\vec{0}\}$

מתקיים:
$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \Rightarrow A^{-1} = \frac{1}{\det(A)}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

מטריצת סיבוב: מטריצה אורתוגונלית ($AA^T = A^TA = I$) עבורה $\det(A) = 1$.
$$A = \begin{pmatrix} \cos(\Theta) & -\sin(\Theta) \\ \sin(\Theta) & \cos(\Theta) \end{pmatrix}$$

הטלה: $P_u(v) = \|v\|\cos(\Theta)\cdot\frac{u}{\|u\|} = \frac{\langle v,u\rangle}{\|u\|^2}\cdot u$

<u>הסתברות</u>

שערוך תוחלת: $\hat{\mu} = \frac{1}{m}\sum_{i=1}^m x_i$
שערוך שונות: $\hat{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^m (x_i - \hat{\mu})^2$
שערוך מטריצת Covariance:
$$\hat{C} = \frac{1}{m-1}\sum_{i=1}^{m}(x_i-\bar{x})(x_i-\bar{x})^T$$

מתאם פירסון: בין X,Y מ"מ:
$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

אי-שוויון מרקוב: עבור X מ"מ אי-שלילי: $\Pr[X \ge t] \le \frac{\mathbb{E}[X]}{t}$
אי-שוויון צ'בישב: $\Pr[|X - \mathbb{E}[X]| \ge t] \le \frac{Var[X]}{t^2}$
אי-שוויון הופדינג: עבור $X_1,\dots,X_n$ ב"ת כך ש-$a_i \le X_i \le b_i$.
$$\Pr[|X - \mathbb{E}[X]| \ge t] \le 2e^{-\frac{2m^2t^2}{\sum_{i=1}^m (b_i-a_i)^2}}$$

טענה: עבור מטריצה A (קבועה), ווקטור מקרי $\boldsymbol{z}$ מתקיים:
$$Var[A\boldsymbol{z}] = A\cdot Var[\boldsymbol{z}]\cdot A^T$$

<u>אינפי</u>

גרדיאנט: $\nabla f = \left(\frac{\partial f}{\partial x_1},\dots,\frac{\partial f}{\partial x_n}\right)$

יעקוביאן:
$$J_x(f) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix} = \begin{pmatrix} | & & | \\ \nabla f_1 & \cdots & \nabla f_n \\ | & & | \end{pmatrix}$$

כלל השרשרת: $J_x(g\circ f) = J_{f(x)}(g)\cdot J_x(f)$

הסיאן:
$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial^2 x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial^2 x_n} \end{pmatrix}$$

נגזרות מפורסמות:
$$\nabla\left(||w||^2\right) = 2w, \qquad J_w(A\cdot w) = A, \qquad J_x(f) = (\nabla f)^T$$