רקע מתמטי

אלגברה לינארית

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:
A מדרגה מלאה, $\det(A) \neq 0, im(A) = \mathbb{R}^m, \ker(A) = \{\vec{0}\}$

מטריצת הטלה אורתוגונלית סימטרית , $P = P^2, (I-P)P = 0$, $Px = x$ לכל $x \in V$, $||x - v|| \geq ||x - Px||$ לכל $u \in V$.

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

הופכית 2X2: $A = \begin{bmatrix} a & b \\ c & d\end{bmatrix} \Rightarrow A^{-1} = \frac{1}{\det(A)}\begin{bmatrix} d & -b \\ -c & a\end{bmatrix}$

מטריצת סיבוב: מטריצה אורתוגונלית כך ש-$\det(A) = 1$.
$$A = \begin{pmatrix}\cos(\Theta) & -\sin(\Theta)\\ \sin(\Theta) & \cos(\Theta)\end{pmatrix}$$

EVD: עבור A ריבועית וסימטרית, $A = UDU^T$ עם U אורתו' ו-D אלכסונית עם ע"ע של A. $A^k = UD^kU^T$ ובפרט עם k=-1.

SVD: $A = U\Sigma V^T = \begin{bmatrix} | & & | \\ u_1 & \dots & u_m \\ | & & | \end{bmatrix}\begin{bmatrix}\sigma_1 & & 0 \\ & ..\sigma_r & \\ 0 & & 0\end{bmatrix}\begin{bmatrix} - & v_1 & - \\ & \vdots & \\ - & v_d & -\end{bmatrix}$

1. נחשב את A^TA. 2. נמצא את הע"ע ואת הו"ע שלה ונסדר אותם בסדר יורד. 3. נגדיר את השורשים של הע"ע להיות האלכסון של Σ ואת הו"ע להיות העמודות של U. 4. נחשב את AA^T. 5. נמצא את הע"ע ואת הו"ע שלה ונסדר אותם בסדר יורד. 5. נגדיר את הו"ע להיות השורות של V.

מטריצה PSD: הגדרות שקולות:
1. לכל $x \in \mathbb{R}^d$ מתקיים $x^TAx \geq 0$.
2. כל הערכים העצמיים של A הם אי-שליליים.
3. קיימת פונקציה B כלשהי כך ש- $A = B^TB$.

הסתברות

אומד לתוחלת: $\hat{\mu} = \frac{1}{m}\sum_{i=1}^m x_i$,

אומד לשונות: $\hat{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^m (x_i - \hat{\mu})^2$

אומד למטריצת 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: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$

אי-שוויון צ'בישב: $\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{Var[X]}{t^2}$

אי-שוויון הופדינג: עבור $X_1,\dots,X_n$ ב"ת כך ש-$a_i \leq X_i \leq b_i$.
$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{2m^2t^2}{\sum_{i=1}^m (b_i - a_i)^2}}$$

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

אינפי

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

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: $\nabla\left(\frac{1}{2}||Ax - y||^2\right) = A^T(Ax - y)$
$\nabla\left(||w||^2\right) = 2w, \qquad J_w(A\cdot w) = A$

קמירות

קבוצה קמורה: יהי V מרחב וקטורי, קבוצה $C \subset V$ נקראת קמורה אם לכל שני וקטורים $u,v \in C$ ולכל $\alpha \in [0,1]$ מתקיים $\alpha u + (1-\alpha)v \in C$. דוגמאות לקמורות – ת"מ וקטורי, גדור יחידה, *halfspace*, חיתוך / כפל בסקלר של קמורות.

פונקציה קמורה: יהי V מ"ו, C קבוצה קמורה. $f: C \to \mathbb{R}$ נקראת קמורה אם לכל $u,v \in C$ ולכל $\lambda \in [0,1]$ מתקיים $f(\lambda u + (1-\lambda)v) \leq \lambda f(u) + (1-\lambda)f(v)$. דוגמאות לקמורות – נורמה, העתקה אפינית, סכום של קמורות (עד כדי כפל בסקלר).

משפט: אם f קמורה, כל מינימום מקומי הוא גלובלי.

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

סדר ראשון: דיפרנציאבילית היא קמורה אמם לכל x,y מתקיים $f(y) \geq f(x) + \nabla f(x)^T(y - x)$. **סדר שני:** דיפרנציאבלית פעמיים היא קמורה אמם ההסיאן שלה הוא *PSD*.

תכונות השומרות על קמירות:
1. **סכום חיובי** - עבור $f_1,\dots,f_n$ קמורות, $\alpha_1,\dots,\alpha_n \geq 0$, 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).

כללי

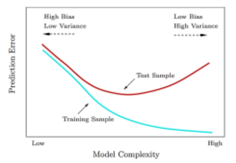
מטריצת אימון: $X \in \mathbb{R}^{m\times d}$ ששורותיה הן הדגימות, ועמודותיה הן התכונות. **וקטור תיוג:** $y \in \mathbb{R}^m$ בעל ערך $y_i = f(x_i) + z_i$ עבור z_i רעש כלשהו.

ההבדלים בשגיאות:
- **Training error / Empirical error:** $L_S(h)$
- **Test error / Generalization error:** $L_{\mathcal{D}}(h)$
- **Approximation error:** $\min_h L_{\mathcal{D}}(h)$ (≈bias)
- **Estimated error:** $L_{\mathcal{D}}(h) - \min_{h'} L_{\mathcal{D}}(h')$ (≈variance)

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	$\frac{1}{m}\sum_{i=1}^m 1_{[h(x)\neq y]}$

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 (נדבק לרעש). ככל שקבוצת ההיפותזות גדולה יותר, כך השונות גדלה. נובעת מ-Estimation Error.

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

רגרסיה

רגרסיה לינארית - מחלקת ההיפותזות:
$$\mathcal{H}_{reg} = \left\{h(x)\middle|h(x) = \langle w, \binom{1}{x}\rangle, w \in \mathbb{R}^{d+1}\right\}$$
(w_0 נקרא intercept. $w_1,\dots,w_d$ נקראים המשקולות)
את המקרה ללא הרעש נפתור עם משוואות נורמליות, את המקרה עם הרעש נפתור עם Maximum Likelihood.

ללא רעש - משוואות נורמליות: מזעור ה-$RSS =$ $argmin_w||y - Xw||^2$ שקול ל- $X^Ty = X^TXw$.

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

פתרון (נוסחא סגורה - אם $y \in ImX$): $\hat{w} = (X^TX)^{-1}X^Ty$.

פתרון (נכון תמיד): $\hat{w} = X^\dagger y$
$$X^{\dagger,\epsilon} = V\Sigma^{\dagger,\epsilon}U^T \;\; s.t\; [\Sigma^{\dagger,\epsilon}]_{ii} = \begin{cases}\frac{1}{\sigma_i} & \sigma_i > \epsilon \\ 0 & otherwise\end{cases}$$

טענות מההוכחה: 1. $\mathrm{Ker}(X) = \mathrm{Ker}(X^TX)$. **2.** עבור מטריצה A ריבועית, $Im(A^T) = Ker(A)^\perp$. **3.** תהי $y = Xw$ מערכת לא הומוגנית, ונניח X לא הפיך. למערכת אינסוף פתרונות אם"ם $y \perp Ker(X^T)$ **4.** למערכת $X^TXw = X^Ty$ יש אינסוף פתרונות או פתרון יחיד (אם X^TX הפיכה).

עם רעש - 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:
$$L(\Theta) = L(\Theta|x_1,\dots,x_m) = \Pr_{\mathcal{D}}[x_1,\dots,x_m|\Theta]$$

(Maximum Likelihood Estimator) 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:
$$p(y|w) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{\left(x_i{}^Tw - y_i\right)^2}{2\sigma^2}}$$
$$= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^m}\prod_{i=1}^m e^{-\frac{\left(x_i{}^Tw - y_i\right)^2}{2\sigma^2}} = L(w|y)$$

פתרון הרגרסיה באמצעות אומד נראות מקסימלית:
$$\widehat{w_{MLE}} = \arg\max_w L(w|y) \overset{by\ log}{=} \arg\min_w \sum_{i=1}^m (x_i{}^Tw - y_i)^2$$
הראינו ששימוש ב-MLE משיג תוצאה זהה ל-Least Squares.

Bias & Variance - Linear Regression			
Low Bias	$(d\sim m)$	$(d \ll m)$	High Bias
Low Var	$(d \ll m)$	$(d\sim m)$	High Var

התאמה פולינומית- מחלקת ההיפותזות:
$$\mathcal{H}_{poly} = \{x \to p_w(x)|\ w \in \mathbb{R}^{d+1}\}$$
כלומר נמצא את $w = \text{argmin}_{w\in\mathbb{R}^{d+1}}\frac{1}{m}\left(y_i - p_w(a_i)\right)^2$, וקטור המקדמים, תוך שימוש ב-LS (אינטואיטיבית – למצוא פולימום שהמרחק שלו מכל הנקודות בדאטא הוא מינימלי).

איך מחשבים: ניקח פיצ'ר אחד (עמודה $(x_1,..,x_m$) וניצור ממנה מטריצת ונדרמונדה: $X = \begin{pmatrix}1 & x_1^1 & \cdots & x_1^d \\ \vdots & & \ddots & \vdots \\ 1 & x_m^1 & \cdots & x_m^d\end{pmatrix}$, ואז פותרים רגרסיה לינארית (עם הפתרון הסגור למשוואות הנורמליות).

קלסיפיקציה

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

Perceptron Algorithm:

Initialize: $w = \vec{0}$
While $\exists i$ s.t $y_i\cdot\langle x_i, w\rangle \leq 0$:
$\quad w \leftarrow w + y_i\cdot x_i$

SVM מחלקת ההיפותזות $\mathcal{H}_{SVM} = \{x \mapsto sign(\langle x,w\rangle + b)\}$

מרחק: $D\left(\binom{w}{b}, x_i\right) = \min_{v:\langle v,w\rangle + b = 0}\|x_i - v\|$

אם $\|w\| = 1$ אז $D\left(\binom{w}{b}, x\right) = |\langle x,w\rangle + b|$

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

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

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 $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(h)$ כלומר המסווג בייס נותן סיכון אמפירי אופטימלי.

LDA: שיטה בה אנחנו מניחים שהדאטא נדגם משתי התפלגויות נורמליות <u>עם אותה שונות Σ</u> ותוחלות שונות μ_0,μ_1, כש-y מתפלג ברנולי עם p כלשהו. מחזיר מפריד לינארי.
עבור $\mathcal{X} = \mathbb{R}^d$ והתפלגות $x|y \sim \mathcal{N}(\mu_y, \Sigma)$, כלומר:
$$f(x|y) = \frac{1}{\sqrt{(2\pi)^d\det(\Sigma)}}e^{-\frac{1}{2}(x - \mu_y)^T\Sigma^{-1}(x - \mu_y)}$$

חישוב באמצעות:
$$h_{\mathcal{D}}(x) = \arg\max_y \underbrace{x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y{}^T\Sigma^{-1}\mu_y + \ln(\Pr[y])}_{\delta_y(x)}$$
כאשר נשערך ע"י: $Pr[y] = \frac{1}{m}\sum_{i=1}^m 1_{[y_i = y]}$, $\mu_y = \frac{\sum_{i=1}^m x_i 1_{[y_i=y]}}{\sum_{i=1}^m 1_{[y_i=y]}}$
וגם: $\Sigma = \frac{1}{m}\sum_{y\in\{\mp1\}}\sum_{i\in[m]:y_i=y}(x_i - \mu_y)(x_i - \mu_y)^T$

QDA: שיטה בה אנחנו מניחים שהדאטא נדגם משתי התפלגויות נורמליות <u>עם שונויות שונות Σ_0,Σ_1</u> ותוחלות שונות μ_0,μ_1 כש-y מתפלג ברנולי עם p כלשהו. מחזיר מפריד ריבועי (פרבולה).

Logistic Regression
מודל הסתברותי עם רעש, משתמש באותו עקרון למידה (נראות מקסימלית) כמו המקרה המורעש של רגרסיה לינארית אבל פועל בדרך אחרת. מודל מאוד interpretable.

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 $w \in \mathbb{R}^{d+1}$.

חישוב: נשתמש ב-logit/סיגמואיד: $\pi(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$
$$\mathcal{H}_{logistic}^d = \{x \mapsto \pi(\langle x,w\rangle)\}$$
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.

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

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

חיתוך (cutoff): המודל מחזיר ערך רציף בין 0 ל-1. אם אנחנו מעוניינים בקלסיפיקציה נבצע-
$$\hat{y} = \begin{cases}1 & h_w(x) \geq \alpha \\ 0 & h_w(x) < \alpha\end{cases}$$

נבחר את α בעזרת **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.

KNN - K Nearest Neighbors
נגדיר מרחק ממושקל $\rho(x_i,z) = \sum_{j=1}^d w_j\cdot\left(x_j - y_j\right)^2$, ועבור k השכנים הקרובים נבחר: $\arg\max_{y\in\mathcal{Y}}\sum_{i=1}^k 1_{\{y_{\pi(i)} = y\}}$

מציאת השכנים ע"י: לכל דגימה, מחשבים את המרחק מכל הדגימות האחרות, ממיינים לפי הגודל, ובוחרים את k הקרובים. בחירת k מתבצעת לפי איזון – נבחר נקודה לפני שהשונות עולה.

Bias & Variance - KNN		
Low Bias	Small $\overset{k}{\to}$ Big	High Bias
Low Var	Big $\overset{k}{\leftarrow}$ Small	High Var

Classification Trees – עצי החלטה
$$\mathcal{H}_{CT}^k = \left\{x \mapsto \sum_{j=1}^N c_j\cdot 1_{\{x_i\in B_j\}}\right\}$$
N כמות הקופסאות, k עומק מקסימלי. $\mathbb{R}^d = \biguplus_j B_j$.
פונקציית התועלת: $P_y^S(B) = \frac{1}{n_s(B)}\sum_{x_i\in B}1_{\{y_i = y\}}$, לפי עיקרון ה-ERM מי שממקסם את ההסתברות זה כלל הרוב הפשוט.

גידול עץ: נחלק לתיבות B_j, בכל שלב נבחר פיצ'ר d וספליט m שממזער את $L_s(T) = \sum_{j=1}^N\sum_{i:x\in B}\left(y_i - \widehat{y_s}(B_j)\right)$. נמשיך עד להגעה לעומק עץ מקס' / כמות מינ' של נקודות בקופסא.

זמן ריצה: $O(md\cdot 2^k) \overset{2^k\leq m}{=} O(m^2d)$.

גיזום העץ: נגדל עץ לגודל מקסימלי ונגזום (נאחד קופסאות שהפיצול שלהן מיותר). הגיזום מוריד variance. בכל שלב נגזום רק אם לאחר הגיזום מזערנו את $\min_{T\subset T_0}[L_S(T) + \lambda|T|]$
(רגולריזציה עם $|T|$ מס' העלים/קופסאות).

Bias & Variance - Decision Tree		
Low Bias	Deep $\overset{height}{\longleftarrow}$ Shallow	High Bias
Low Var	Shallow $\overset{height}{\longrightarrow}$ Deep	High Var

סיווג שגיאות קלסיפיקציה: שגיאת <u>**Type-I**</u> (או FP) זו השגיאה החמורה. שגיאת <u>**Type-II**</u> (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: $\frac{TN}{TN + FN}$	$F_1 = 2\cdot\frac{Recall\cdot Precision}{Recall + Precision}$

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 הריאלזביליות)
כאשר נריץ את $\mathcal{A}$ על $m \geq \widetilde{m}_{\mathcal{H}}(\varepsilon,\delta)$ 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: $\lim_{m\to\infty}\mathbb{E}_{S\overset{iid}{\sim}\mathcal{D}^m}\left[L_{\mathcal{D},f}(\mathcal{A}(S))\right] = 0$

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

אין ארוחות חינם: יהי $\mathcal{X}$ מרחב דגימות אינסופי, ונקבע $\varepsilon < \frac{1}{2}$. קיימת $\delta > 0$ כך שלכל אלגוריתם למידה $\mathcal{A}$ וקורפוס אימון בגודל m, קיימת התפלגות $\mathcal{D}(x)$ מעל $\mathcal{X}$ ופונקצייה $f: \mathcal{X} \to \mathcal{Y}$ כך ש:
$$\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}\left[L_{\mathcal{D},f}(\mathcal{A}(S)) > \varepsilon\right] \geq \delta$$

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: $\mathcal{H}_C = \{h\restriction_C: h \in \mathcal{H}\}$
$$VCdim(\mathcal{H}) = \max\left\{|C| : C \subset \mathcal{X}\ and\ |\mathcal{H}_C| = 2^{|C|}\right\}$$

טבלת VCdim

Class	VCdim
Axis aligned rectangles 2d	4
Axis aligned rectangles 3d	6
Intersection of 2 Axis alig' rectangles 2d	4
Homogeneous Halfspaces	d
Non- Homogeneous Halfspaces	$d + 1$
Polynomial Thresholds (Pol_r)	$\leq \binom{d+r-1}{d}$
k-interval	$2k$
Conjuctions	d

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

הכללת המושגים:
• התפלגות $\mathcal{D}$ מעל $\mathcal{X}\times\mathcal{Y}$ • Approximately Correct עבור $h \in \mathcal{H}$ המקיימת $L_{\mathcal{D}}(h) \leq \min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$ • 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 $m \geq \widetilde{m}_{\mathcal{H}}(\varepsilon,\delta)$ דגימות $i.i.d$ מעל $\mathcal{D}$, מתקיים: $\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}\left[L_{\mathcal{D},f}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\right] \geq 1 - \delta$

המשפט היסודי של הלמידה הסטטיסטית:
תהי $\mathcal{H} \subset \{\pm1\}^{\mathcal{X}}$ מחלקת היפותזות ונסמן $d = VCdim(\mathcal{H})$. אז $\mathcal{H}$ למידה PAC אם"ם $\mathcal{H}$ למידה Agnostic-PAC אם"ם $d < \infty$. במקרה זה:
• קיימים קבועים אוניברסליים C_1,C_2 כך ש-$\mathcal{H}$ למידה PAC עם סיבוכיות דגימות:
$$C_1\frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon,\delta) \leq C_2\frac{d\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)}{\varepsilon}$$
• קיימים קבועים אוניברסליים C_1,C_2 כך ש-$\mathcal{H}$ למידה Agnostic-PAC עם סיבוכיות דגימות:
$$C_1\frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon,\delta) \leq C_2\frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$$
• 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 $L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$

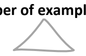
התכנסות אחידה: למחלקת היפותזות $\mathcal{H}$ יש תכונת התכנסות אחידה אם קיימת $m_{\mathcal{H}}^{UC}: (0,1)^2 \to \mathbb{N}$ כך שלכל $\varepsilon,\delta \in (0,1)$ ולכל התפלגות $\mathcal{D}$ מעל $\mathcal{X}\times\mathcal{Y}$ מתקיים:
$$\Pr_{S\overset{iid}{\sim}\mathcal{D}^m}[S\ is\ \varepsilon - representative] \geq 1 - \delta$$

טענה: אם $\mathcal{H}$ בעלת תכונת התכנסות אחידה עם פונקציה $m_{\mathcal{H}}^{UC}: (0,1)^2 \to \mathbb{N}$ אז $\mathcal{H}$ למידה Agnostic-PAC עם סיבוכיות דגימות $m_{\mathcal{H}}(\varepsilon,\delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2},\delta\right)$.

מחלקות סופיות

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

Number of examples (m)



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

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

שיטות ועידה Ensemble Methods

טענה: בהינתן מ"מ $X_1,\dots,X_T$ בעלי שונות σ^2 וקורלציה $corr(X_i,X_j) = \rho$. השונות של $\bar{X} = \frac{1}{T}\sum_{t=1}^T X_t$ 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 $h(x) = \frac{1}{T}\sum_t h_t(x)$

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!

Bootstrap
בהינתן קבוצת דגימות S אנו יוצרים B קבוצות חדשות ע"י דגימה חוזרת מתוך S במשך m פעמים (הראינו שזה מדמה בצורה טובה דגימה מ-D המקורי ש-S נדגמה מתוכו):
$$S^b = \{(x_i^b, y_i^b)\}_{i=1}^m$$

Empirical Distribution: התפלגות $\widehat{\mathcal{D}}_S$ המושרת ע"י S מעל $\mathcal{X}\times\mathcal{Y}$ על תת-קבוצה $C \subset \mathcal{X}\times\mathcal{Y}$:
$$\widehat{\mathcal{D}}_S\left((X,Y) = (x,y)\right) = \begin{cases}\frac{1}{m} & (x,y) \in S \\ 0 & otherwise\end{cases},\ \ \widehat{\mathcal{D}}_S(C) = \frac{|C\cap S|}{m}$$

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 CDF (הצטברות משותפת) היא פונקציית מדרגות העולה מ-0 ל-1 בקפיצות של $\frac{1}{m}$ 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:
$$h_{bag}(x) = sign\left(\sum_{t=1}^T h_{S^t}(x)\right)$$

דה-קורלציה: מבוצעת כיוון שניתן להוריד את השונות רק עד מידה מסוימת. למשל בעצים נבחר פרמטר k ובכל עת נבחר k פרמטרים שרק מתוכם ניתן לפצל.

Bias & Variance - Bagging		
Bias	Doesn't change	**Bias**
Low Var	$\text{Big}(\infty) \xleftarrow{T} \text{Small}$	**High Var**

Random Forest שימוש ב-*Bagging* כדי לשפר את הביצועים של עצי החלטה.

Random Forest Algorithm:

- מאתחלים T עצים בעומק R באמצעות m_{min} דגימות
- לכל עץ $t\in[1,T]$:
  - מייצרים דגימת *bootstrap* - S^* מתוך S
  - מאמנים עץ החלטה עם הדגימה ומקבלים היפותזה h_{S^*}
  - כל עוד לא הגענו לעומק מקסימלי או מספר מינימלי של נקודות בקוביה:
    - בוחרים בצורה אחידה k פיצ'רים מתוך d הפיצ'רים
    - בוחרים את הפיצ'ר הטוב מביניהם ומבצעים הפרדה לפיו
- נחזיר את $h_S(x)=sign(\sum_{t=1}^{T}w_t h_t(x))$

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 זמין אך יכול ליצור כפילויות דגימה]

Adaboost Algorithm:

Initialize: $D^1=\left(\frac{1}{m},\dots,\frac{1}{m}\right)$
Loop: for t in [T]: $h_t=\mathcal{A}(D^t,S)$
Update:
- $\sum_{j=1}^{m}D_j^{t+1}\cdot 1_{[h_t(x_j)\neq y_j]}=\frac{1}{2}\Rightarrow w_t=\frac{1}{2}\log\left(\frac{1}{\varepsilon_t}-1\right)$
  ($\varepsilon_t=\sum_{j=1}^{m}D_j^{t}\cdot 1_{[h_t(x_j)\neq y_j]}$,סיכון אמפירי ממושקל)
- $D_i^{t+1}=\frac{D_i^t\cdot e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^{m}D_j^t\cdot e^{-w_t y_j h_t(x_j)}}$

<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).
חיזוי דגימה חדשה:
$$h_{boost}(x)=sign\left(\sum_{t=1}^{T}w_t h_t(x)\right)$$

Bias & Variance - Adaboost		
Low Bias	$\text{Big}(\infty) \xleftarrow{T} \text{Small}$	**High Bias**
Low Var	$\text{Small} \xrightarrow{T} \text{Big}(\infty)$	**Var goes up a bit**

לומדים חלשים
γ-weak-learner: אלגוריתם ללמידה $\mathcal{A}$ 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 $f:\mathcal{X}\to\{\pm1\}$, אם הנחת הריאלזביליות 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:
$$\Pr_{S\sim_{iid}\mathcal{D}^m}\left[L_{\mathcal{D},f}(h_S)\le\frac{1}{2}-\gamma\right]\ge1-\delta$$
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) עבורו הסיכון האמפירי הממושקל מקיים
$$\varepsilon_t=\sum_{j=1}^{m}D_j^t\cdot 1_{[h_t(x_j)\neq y_j]}\le\frac{1}{2}-\gamma$$
אז כלל הסיכון האמפירי (הלא ממושקל) של כלל ההחלטה של Adaboost מקיים:
$$L_S(h_{boost})\equiv\frac{1}{m}\sum_{j=1}^{m}1_{[h_{boost}(x_j)\neq y_j]}\le e^{-2\gamma^2T}$$

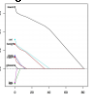
	Bagging	***Boosting***
מקביליות	במקביל	טורי
מבנה הנתונים	*Bootstrap training samples*	*Bootstrap* ממושקל או קבוצה S המקורית עם כלל *ERM* ממושקל
De-correlation	מומלץ	אין צורך
Overfitting	לא קורה	סיכון
איזה מודל להשתמש	עצים עמוקים	עצים רדודים
השפעה	מפחית *Variance*	מפחית *Bias*

רגולריזציה
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 הרגולציה.
- 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 לפונקציה המטרה

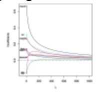
Bias & Variance - Regularization		
Low Bias	$0 \xrightarrow{\lambda} \infty$	**High Bias**
Low Var	$\infty \xleftarrow{\lambda} 0$	**High Var**

רגולריזציה של רגרסיה לינארית
- **Ridge Regression**
$$\arg\min_{w_0\in\mathbb{R},w\in\mathbb{R}^d}\left\|w_0\cdot\vec{1}+Xw-y\right\|^2+\lambda\cdot\|w\|_2^2$$
יש לפתור את המערכת $X^Ty=(X^TX+\lambda I)w$. מתקיים $\hat{w}_\lambda=U\Sigma^\lambda V^Ty$ כאשר $\Sigma^\lambda=diag\left(\frac{\sigma_i}{\sigma_i^2+\lambda}\right)$ או-$w=(X^TX+\lambda I)^{-1}X^Ty$ עבור $\lambda>0$
- **Lasso** – בעל תכונת דלילות.
$$\arg\min_{w_0\in\mathbb{R},w\in\mathbb{R}^d}\left\|w_0\cdot\vec{1}+Xw-y\right\|^2+\lambda\cdot\|w\|_1$$
- **Best Subset**
$$\arg\min_{w_0\in\mathbb{R},w\in\mathbb{R}^d}\left\|w_0\cdot\vec{1}+Xw-y\right\|^2+\lambda\cdot\|w\|_0$$

Lasso Regularization Path



Ridge Regularization Path



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

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.
$$\sigma=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$
Bootstrap: נדגום B קבוצות S^b, ונאמן את האלגוריתם עליה. בכדי לבחון את הביצועים נשתמש ב-$T^b=S\setminus S^b$ ונמצע במקודם.

טעויות בבחירת מודל
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 עיבוד המתאים למודל.

השגיאה בעת שימוש בולידציה
$$L_{\mathcal{D}}(h_S)=\underbrace{L_{\mathcal{D}}(h_S)-L_{\mathcal{V}}(h_S)}_{A}+\underbrace{L_{\mathcal{V}}(h_S)-L_S(h_S)}_{B}+\underbrace{L_S(h_S)}_{C}$$
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 $\delta\in(0,1)$:
$$\Pr\left[|L_{\mathcal{V}}(h)-L_{\mathcal{D}}(h)|\le\sqrt{\frac{\ln(2/\delta)}{m_{\mathcal{V}}}}\right]\ge1-\delta$$
עבור $\mathcal{H}_1\subseteq\dots\subseteq\mathcal{H}_k$ סופיים.
Standard Method: עבור $h^*\in ERM_{\mathcal{H}_k}(S_{all})$:
$$\Pr\left[L_{\mathcal{D}}(h^*)\le\min_{h\in\mathcal{H}_k}L_{\mathcal{D}}(h)+\sqrt{\frac{2\ln(2|\mathcal{H}_k|/\delta)}{m}}\right]\ge1-\delta$$
Model Selection: עבור $h^*\in\arg\min_{h\in\mathcal{H}_k}L_{\mathcal{D}}(h)\subseteq\mathcal{H}_j\subseteq\dots\subseteq\mathcal{H}_k$
$$\Pr\left[L_{\mathcal{D}}(h^*)\le\min_{h\in\mathcal{H}_k}L_{\mathcal{D}}(h)+\sqrt{\frac{2}{\alpha m}\ln\left(\frac{4k}{\delta}\right)}+\sqrt{\frac{2}{(1-\alpha)m}\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}\right]\ge1-\delta$$

Unsupervised Learning
דוגמאות לשימוש: 1. חשיפה של מבנה במימד נמוך. **2.** קלסיפיקציה. **3.** זיהוי אנומליות.

PCA
$$\arg\min_{U\in\mathbb{R}^{d\times k},W\in\mathbb{R}^{k\times d}}\sum_{i=1}^{m}\|x_i-UWx_i\|^2$$
משפט: תהי $A=\sum_{i=1}^{m}x_ix_i^T$, ויהיו $u_1,\dots,u_n$ הו"ע של A (מסודרים בסדר יורד), אז $W=U^T=[u_1\cdots u_k]$ פתרון לבעיית ה-PCA.
- $P=\sum_{i=1}^{k}u_iu_i^T$ נקראת מטריצת הטלה, היא סימטרית והיא קירוב טוב ביותר ל-x בתת-מרחב:
$$\|x-v\|_2\ge\|x-Px\|_2\ \ \forall v\in span\{u_1,\dots,u_k\}$$
- 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 $\lambda_1\ge\dots\ge\lambda_d\ge0$.
- 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: לוקחים את $A=\sum_{i=1}^{m}(x-\bar{x})(x-\bar{x})^{\top}$ (מטריצת ה-*Covariance* של X), מחשבים את הו"ע שלה ולוקחים את k הגדולים מביניהם (וקטורים מובילים), מדביקים אותם כעמודות למטריצה U. כעת $U^{\top}X$ פותרת את *PCA* – הטלה לתת המרחב הנכון ממימד k (אלו **נקודות אינטרינזיות** – נקודות בתוך תת המרחב <u>אליו הטלנו</u> ממימד k), ו- $UU^{\top}X$ נותנת את הקואורדינטות של הנקודה לאחר ההטלה אבל <u>במרחב המקורי</u> (ממימד d).
לאחר מכן כל וקטור במרחב ניתן לייצוג ע"י הקירוב $\bar{x}+\sum\alpha_iu_i$.
אם יש יותר דגימות מפיצ'רים, חישוב A הוא כבד $O(d^3)$ ולכן נרצה להשתמש באלגוריתם הבא:

PCA Algorithm:

Input: $X\in\mathbb{R}^{m\times d}$
Eval:
- if $m>d$:
  - $A=X^TX$
  - $u_1,\dots,u_k$ eigenvectors of A
- else ($m\le d$):
  - $B=XX^T$
  - $v_1,\dots,v_k$ eigenvectors of B
  - denote $u_i=\frac{1}{\|X^Tv_i\|}X^Tv_i$
- return $u_1,\dots,u_k$

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

Clustering
נגדיר חלוקה של המידע $x_1,\dots,x_m$ ל-k מחלקות $\sqcup_jC_j$. 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: הערך (של C_j) $\arg\min_{\mu_j}\sum_{x\in C_j}d(x,\mu_j)^2$.
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: $x_1,\dots x_m$ and $k\in\mathbb{N}$
Step 0: choose initial $\mu_1,\dots,\mu_k$
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: $\mu_j=\frac{1}{|C_j|}\sum_{x\in C_j}x$

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).

Bias & Variance - Clustering		
Low Bias	$\text{Big} \xleftarrow{k} \text{Small}$	**High Bias**
Low Var	$\text{Small} \xrightarrow{k} \text{Big}$	**High Var**

Spectral Clustering
אלגוריתם שמבצע צעדים מקדימים חכמים לפני שהוא מפעיל *k-means*, נרצה להתעלם ממרחקים גדולים מידי, כלומר נחשוב עליהם בתור אינסוף ונתחשב רק במרחקים קטנים. ואז נבנה גרף סימטרי וממושקל שהקודקודים שלו הם mהדגימות שלנו, ותהיה קשת בין שני קודקודים אם"ם המרחק ביניהם קטן (כך תבוא לידי ביטוי ההתעלמות ממרחקים גדולים) נביט ברכיבי הקשירות של הגרף ולפיו נקבע את הקלאסטרים.
נגדיר את מטריצת הסמיכויות $A_{i,j}=exp\left(-\frac{\|x_i-x_j\|^2}{\varepsilon}\right)$ (שמגלמת שככל שהדגימות יותר דומות האפיניטי שלהן יותר גדול, אם הוא 1 הן זהות). נמצא את k הו"ע של $L=D^{-1}A$ (שנקראת גרף לפלסיאן), כאשר $D=\sum_{j=1}^{m}A_{ij}$, שמתאימים לע"ע 1. הם ייצגו את הקלסטרים, וכל i תיוצג במימד k ע"י הקואורדינטות ה-i שלהם. כעת נפעיל $k-means$ במימד k.

Kernel methods
סוג של בעיה הפוכה מ- *PCA* - אנחנו רוצים להטיל למרחב <u>גבוה</u> יותר כדי להפיק מידע נוסף מהדגימות.
נפתור בעיות אופטימיזציה מהצורה:
$$w^*=\arg\min_{w\in\mathbb{R}^k}\|f(\langle w|\psi(x_1)\rangle,..,\langle w|\psi(x_m)\rangle)\|^2+R(\|w\|_2^2)$$
לפי משפט הייצוג, קיים $\alpha\in\mathbb{R}^m$ כך ש- $w^*=\sum_{i=1}^{m}\alpha_i\psi(x_i)$.
תהי $G\in\mathbb{R}^m$ המוגדרת ע"י: $G_{i,j}=\langle\psi(x_i)|\psi(x_j)\rangle$. הבעיה הנ"ל שקולה ל: $\arg\min_{\alpha\in\mathbb{R}^m}f(G\alpha)+R(\alpha^TG\alpha)$. והחיזוי מתבצע ע"י: $y(x)=\langle w^*|\psi(x)\rangle=\alpha^Tk$, כאשר $k_i=\langle\psi(x_i)|\psi(x)\rangle$.
פונקציה סימטרית K תיקרא ***<u>PSD-kernel</u>*** אמ"ם לכל $m\in\mathbb{N}$ ולכל $x_1,\dots,x_m\in\mathcal{X}$ המטריצה $K_{i,j}=k(x_i,x_j)$ היא PSD.
לכן ע"מ להראות שהעתקה k כלשהי היא פונקציית קרנל נראה 1. מטריצת גראם המזוהה איתה היא PSD. או 2. קיימת פונקציה ψ עבורה היא ממששת מכפלה פנימית במרחב החדש.
תנאי מרסר: תהי $\psi:\mathcal{X}\to\mathcal{F}$ כאשר $\mathcal{F}$ מרחב הילברט, אזי פונקציה $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ מממשת בו מכפלה פנימית אמ"ם היא PSD-kernel.
יצירת קרנלים חדשים:
יהיו $K_1(x,x'),K_2(x,x')$ פונקציות קרנל חוקיות, כל הבאים הם קרנלים חוקיים:
1. $K(x,x')=f(x)K_1(x,x')f(x')$ לכל פונקציה f.
2. $K(x,x')=\alpha K_1(x,x')+\beta K_2(x,x')$ עבור $\alpha,\beta\ge0$.
3. $K(x,x')=K_1(x,x')K_2(x,x')$.

דוגמאות לקרנלים:
בעיית ה ridge regression מקיימת: $y(x)=k^T(G+\lambda I)^{-1}y$.
פונקציית קרנל עבור התאמת פולינומיאלים מדרגה לכל היותר k מתקבלת ע"י: $k(x,x')=(1+\langle x|x'\rangle)^k$.

תכונות של תת-גרדיאנט
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 הסאב-גרדיאנטים של f בנקודה w.
למה: f קמורה אם"ם לכל נקודה $w\in dom(f)$ מתקיים $\partial f(w)\neq\emptyset$. 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 היא מינימום גלובלי של f.

ליפשיציות
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-ליפשיץ אם"ם הנורמה של כל סאב-גרדיאנט של f הוא לכל היותר ρ.

אופטימיזציה קמורה ו-Gradient Descent
בעיית אופטימיזציה:
$$\min_x\ f_0(x)\qquad s.t\ \ f_i(x)\le b_i\quad\forall i=1,\dots,n$$
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.
- **חסומה $\mathcal{H}$** – קיים B עבור $\|w\|\le B$ $\forall w\in\mathcal{H}$.
- **ליפשיץ ℓ** – הפונקציה $\ell(h(x),y)$ קמורה ב-h והיא ρ-ליפשיץ.

במקרה זה סיבוכיות הדגימה תלויה רק ב-$\rho,B,\delta,\varepsilon$.

Sub-gradient Descent Algorithm:

initialize: $w^{(1)}=0$
for $t=1,\dots,T$:
- Choose $v_t\in\partial f(w^{(t)})$
- $w^{(t+1)}\leftarrow w^{(t)}-\eta\cdot v_t$

return $\bar{w}=\frac{1}{T}\sum_{t=1}^{T}w^{(t)}$

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

Stochastic Gradient Descent Algorithm:

initialize: $w^{(1)}=0$
for $t=1,\dots,T$:
- Choose $(x,y)\sim\mathcal{D}$
- Choose $v_t\in\partial\ell\left(w^{(t)},(x,y)\right)$
- $w^{(t+1)}\leftarrow w^{(t)}-\eta\cdot v_t$

return $\bar{w}=\frac{1}{T}\sum_{t=1}^{T}w^{(t)}$

למה: תהי בעיית למידה קמורה-ליפשיץ-חסומה עם פרמטרים ρ,B. אז לכל $\varepsilon>0$, אם נריץ את האלגוריתם SGD למזעור $L_{\mathcal{D}}(w)$ עם מספר איטרציות $T\ge\frac{B^2\rho^2}{\varepsilon^2}$ ועם $\eta=\sqrt{\frac{B^2}{\rho^2T}}$, 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.

רשתות נוירונים
פונקציות אקטיבציה מוכרות:
Sigmoid(x) $=\frac{1}{1+e^{-x}}$, **Tanh**(x) $=\frac{2}{1+e^{-2x}}-1$
reLU(x) $=\begin{cases}0 & x<0\\ x & x\ge0\end{cases}$
באלגוריתם SGD של רשתות בוחרים את w ההתחלתי באופן <u>רנדומלי</u> (עם התפלגות כך ש- $w^{(1)}$ קרוב מספיק ל-0).
את הגרדיאנט מחשבים בעזרת backpropagation:

Backpropagation Algorithm:

input: example (x,y), weight vector W, layered graph (V, E) and activation function $\sigma:\mathbb{R}\to\mathbb{R}$
initialize: denote layers of the graph $V_0,\dots,V_T$ where $\boldsymbol{V_t=\{v_{t,1},\dots,v_{t,k_t}\}}$. define $W_{t,i,j}$ *as weight of* $(\boldsymbol{v_{t,j}},\boldsymbol{v_{t+1,i}})$
forward: set $O_0=x$ **and for** $\boldsymbol{t=1,\dots,T}$**:**
set $a_{t,i}=\sum_{j=1}^{k_{t-1}}W_{t-1,i,j}\cdot O_{t-1,j}$ **and** $O_{t,i}=\sigma(a_{t,i})$
backward: set $\delta_T=O_T-y$ **and for** $\boldsymbol{t=T-1,T-2,\dots,1}$**:**
for $i=1,\dots,\boldsymbol{k_t}$ **set** $\delta_{t,i}=\sum_{j=1}^{k_{t+1}}W_{t,j,i}\delta_{t+1,j}\sigma'(a_{t+1,j})$
output: for each edge $(\boldsymbol{v_{t-1,j}},\boldsymbol{v_{t,i}})\in$ E
set the partial derivative to $\delta_{t,i}\sigma'(a_{t,i})O_{t-1,j}$

פונקציות בסיסיות ב-*Numpy*
יצירת מערך:
```
np.array([#desired array]) | np.eye(size) # unit matrix
np.zeros((x, y, z)) # x, y, z are sizes | np.ones((x, y, z)) |
np.arange(start, stop, step) # [start, start + step, …, stop - 1]
```
אינדקסים:
```
array_2D[2:5, :] # select from 3rd row to 5th, all columns
array_2D[2:5, 1:3] # select from 3rd row to 5th, columns 1, 2
array_2D[2:5, [3, 5, 11]] # from 3rd to 5th, specified columns
```
פעולות על מטריצה:
```
A.reshape(x, y, z) # reshape to the given shape
A + 1 | 2 * A # adds / multiplies, element-wise
np.power(A, 3) # element-wise 3rd power.
np.log(A) | np.exp(A) # element-wise log/exp.
A.T # transpose | A > 10 # [12, 5, 30] => [True, False, True]
```
פעולות בין מטריצות:
```
A + B | A * B # element-wise addition | multiplication
A @ C # matrix multiplication
A > B # [12, 5, 1] > [4, 7, 2] => [True, False, False]
np.concatenate((A, B), axis=0) # 0 for rows, 1 for columns
```
פעולות בסטטיסטיקה:
#את כל אחת מהפעולות הבאות ניתן לקבל עבור:
```
The whole array: np.func(arr)
By row: np.func(arr, axis=1) | By column: np.func(arr, axis=0)
---
np.sum(A) | np.sum(A, axis=1) # sums the matrix/the rows
np.max(A) | np.max(A, axis=0) # max of matrix/each column
np.mean(A) | np.mean(A, axis=1) # mean of matrix/rows
np.sort(A) | np.sort(A, axis=0) # sort matrix (flattened) | cols
np.cov([x1, …, xn], [y1, …, yn]) # covariance matrix
```
אלגברה לינארית:
```
np.linalg.inv # returns inverse matrix
np.linalg.svd # returns (U, Sigma (as array), V.T)
np.linalg.eig # returns (eigenvalues [as array], V [eigenvectors])
```