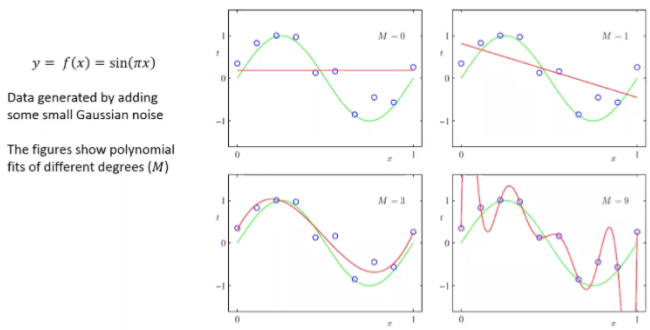
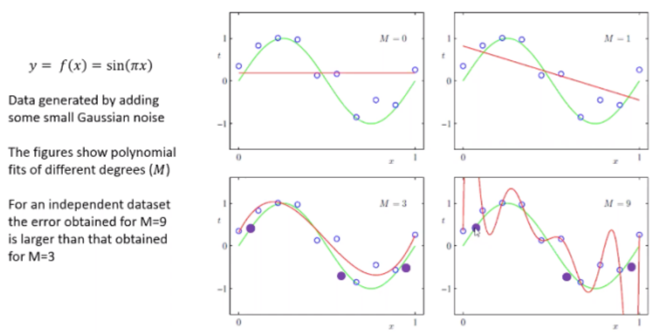
Intro to Overfitting, Param Estimation, Prolog On Prob Learning – 3 למידה חישובית ממידע: הרצאה

Introduction to Overfitting

- **טעות השערוך:** מודד את הטעות בין הפרדיקציה לבין הערך האמיתי. מדדנו טעות זו עד כה על הדאטה אימון. נקראת גם טעות האימון.
- **טעות ההכללה:** מודד את הטעות על דאטה חדש, לא על הדאטה אימון. בלמידה חישובית אנו מעוניינים בטעות ההכללה.

אוברפיטינג ברגרסיה פולינומיאלית:





- נלקחה הפונקציה סינוס(פאי*איקס) ונדגמו ממנה 10 נקודות עליהן חושבה הפונקציה, אלה הן הנקודות הכחולות, אולם זהו לא ערכן המדויק, הוספנו "רעש" גאוסיאני (שמפולג נורמלית) לערך. אם לא היינו מוסיפים רעש הדגימות הכחולות היו נמצאות על הפונקציה הירוקה (זו הפונקציה שאכן יצרה את הדאטה עבורי).
- בעולם האמיתי אנחנו לא יודעים איך נראית הפונקציה שאנו צריכים למצוא – הגרף הירוק הוא לכאורה בלתי נראה.
- ראשית התחלנו בניסוי לחזות את הנקודות על ידי פונקציה ממעלה0 (הקו הישר בגרף הימני למעלה), עשינו זאת גם עבור מעלה 3 וגם עבור מעלה 9.
- בפולינום ממעלה 9 יש 10 משוואות עם 10 נעלמים ולכן ה-MSE (הטעות) הוא 0, אבל הוא מתאים מידי לדאטה שלנו!

- **אם איננו יודעים איך נראה הגרף הירוק איך נדע להחליט ש- m=3 טוב יותר מאשר m=9?**
- **__נבצע ולידציה__:** נגריל 3 נקודות חדשות (הסגולות המלאות באיור משמאל) שלא השתמשנו בהן ב-training ועליהן נמדוד את הטעות, את ה-MSE (באותו האופן שבו הוגרלו 10 הנקודות של האימון).

• Consider the error of a hypothesis/model over:
- The training set data: $error_{train}(h)$
- The entire distribution F of data: $error_F(h)$ ("true error" or "generalization error")

Hypothesis $h \in H$ overfits training data if there is an alternative hypothesis $h' \in H$ such that

$$error_{train}(h) < error_{train}(h')$$

and

$$error_F(h) > error_F(h')$$

הגדרה פורמאלית ל-Overfitting:

נתבונן בשגיאה של ההיפותזה/המודל מעל:

- קבוצת הדאטה אימון: ERROR train (h)
- מעל ההתפלגות F של הדאטה: ERROR F (h) ("הטעות האמתית" או "טעות ההכללה")

היפותזה h מתוך מרחב ההיפותזות H מקיימת overfit על דאטה האימון אם קיימת היפותזה אלטרנטיבית h' מתוך מרחב ההיפותזות H כך שמתקיים

$$error_{train}(h) < error_{train}(h') \quad \text{וגם} \quad error_F(h) > error_F(h').$$

כלומר, הטעות של היפותזה h על הדאטה אימון קטנה מאשר הטעות של היפותזה h' על הדאטה אימון וגם, הטעות של היפותזה h על כל הדאטה גדולה מהטעות של היפותזה h' על כל הדאטה = היפותזה h מותאמת מידי לדאטה אימון!

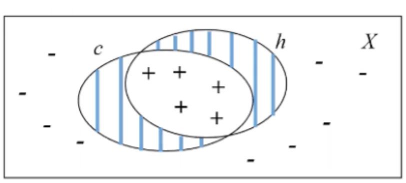
הטעות האמתית (במקרה של סיווג/קלסיפיקציה):

הטעות האמיתית של היפותזה h ביחס לקונספט המטרה c היא ההסתברות ש-h תטעה בסיווג לקלאס של דגימה שנשלפה באופן רנדומלי מתוך ההתפלגות של הדאטה F:

$$error_F(h) = Pr_{x \sim F}[c(x) \neq h(x)]$$

זו למעשה ההסתברות להימצא בשטח המקווקו הכחול באיור המוצג.

השגיאה הזו תלויה מאוד בהתפלגות הדאטה! (ב-F)



<u>הערכה סטטיסטית Statistical Estimation</u>

- נוכל להשתמש ב- test set במטרה להעריך את השגיאה האמיתית של היפותזה מועמדת/מודל מועמד.
- אם ה-test set הוא היתרה של X נדע את השגיאה האמתית! אבל זהו מצב לא ריאליסטי כמובן – אנחנו חייבים להסתמך על דגימות.

נגדיר את טעות הדגימה, עבור סט של דגימות באופן הבא: $error_S(h)$ = the ratio of misclassified samples in S

- ככל שקבוצת הדגימות S גדולה יותר, כך ההערכה תהיה טובה יותר. נרצה להבין את האיכות של ההערכה שלנו.
- הדבר שקול לשאלה הבאה בסטטיסטיקה: העריכו את הקבוצה היחסית של האוכלוסייה (אחוז מהאכלוסייה) בעלת תכונה מסוימת.

במקרה שלנו, התכונה של כל x באוכלוסייה X היא שההיפותזה שלנו h מסווגת את x באופן שגוי.

ההתפלגות של טעות הדגימה – התפלגות בינומית (הצלחה וכישלון של קבוצת ניסויים בלתי תלויים)

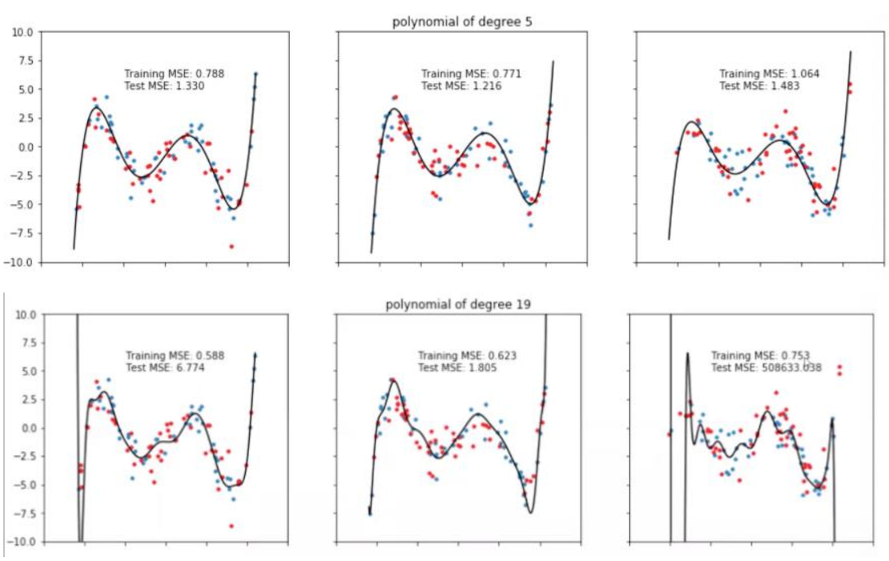
- עבור דגימה ספציפית x, נסמן את ההסתברות למיס-קלסיפיקציה המוגדרת על ידי error F (h) באות p.
- נניח כי קבוצת המדגם שלנו מכילה n דגימות רנדומליות שהוגרלו באופן בלתי תלוי מ-X.
- יהי R משתנה מקרי שמוגדר להיות מספר השגיאות (המיס-קלסיפיקציות) שתניב ההיפותזה h כאשר נפעיל אותה על קבוצת המדגם שלנו.
- אזי: $\text{Prob}(R=k)=\frac{n!}{k!(n-k)!}\cdot p^k(1-p)^{n-k}$
- אבל אנחנו לא יודעים את p! לכן עלינו להשיג הערכה עבור p.

התהליך של ההערכה הסטטיסטית Statistical Estimation Procedure

- נשתמש ב"test set" בגודל |S| = n. ונניח כי מספר הטעויות הוא r.
- נוכל להראות כי r/n הוא הערכה עבור הטעות המוכללת.
- בתלות על הגודל של סט הבדיקה שלנו, נוכל להפיק הבטחה סטטיסטית כגון:

בוודאות של 95% נעריך כי הטעות האמתית קטנה מ- $\frac{r}{n}+\varepsilon$

- תהליך זה נקרא גם: confidence intervals for proportion estimates

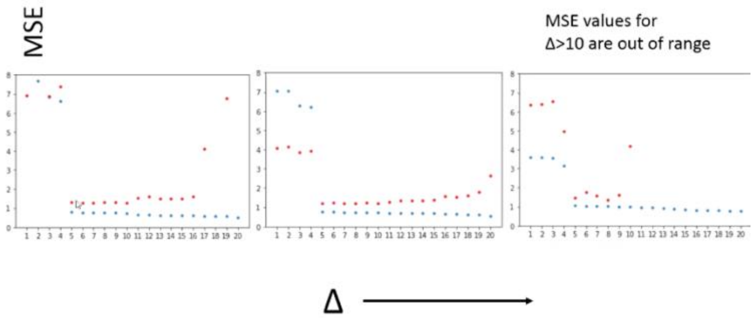


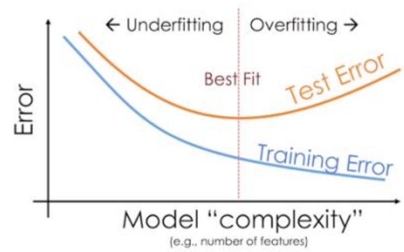
דוגמה נוספת:

יש כאן 50 דגימות אדומות שמהוות טסט ו-50 דגימות training כחולות, של פולינום ממעלה 5 עם רעש.
ה-MSE test מייצג עבורנו את טעות ההכללה (טעות ההכללה האמתית מחושבת על אינסוף גבולות).
ניתן לראות שה- MSE הטוב ביותר הוא כמובן, במעלה 5, נבין את זה רק לאחר שנמשיך למעלות הבאות (להלן מעלה 19 שמניב MSE עם ערך גבוהה מאוד).

נציג את ה-MSE כפונקציה של השינוי במעלה של הפולינומים שבדקנו

ניתן לראות שבמעלות נמוכות 1-4 ה-MSE הוא מאוד גבוה בשני הסטים, במעלה 5 אנחנו כבר יורדים ל-MSE נמוך יותר בשני הסטים. **אבל, במעלות הגבוהות ה-MSE של הtraining set הולך ונעשה נמוך יותר, השגיאה קטנה, ואנחנו מתאימים את המודל שלנו מידי לדאטה, מפני שבמקביל ה-MSE של ה-test set הולך וגדל!**

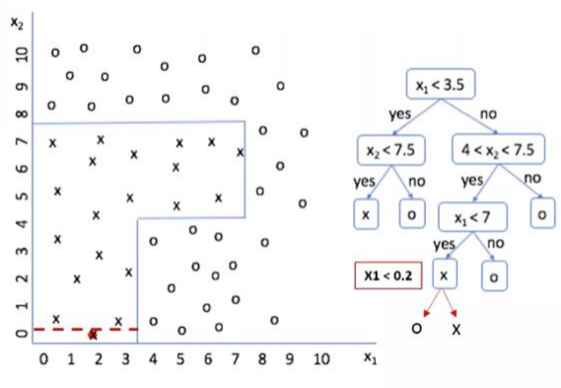




- **Model "complexity"** – בדוגמה לעיל הינו המעלות של הפולינום, ובעצי החלטה המורכבות של המודל מתבטאת בעומק העץ/גובה העץ.
- **Best Fit** – הנקודה בה השגיאה על ה-training set קטנה ועל ה-test set היא מינימלית. בנקודה זו נעדיף לעצור לקדם את מורכבות המודל. נקודה זו תלויה בגודל הדאטה.

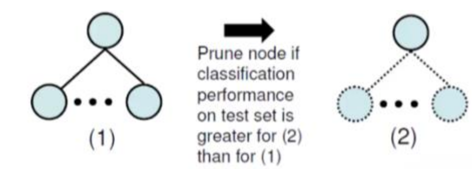
<u>בחזרה לעצי החלטה</u> – נחזור לדוגמה הזו שראינו כבר בהרצאה 2, ונניח כי הייתה לנו טעות בדאטה ומדדנו דגימה שהיא עיגול כאיקס (מסומן באדום) ואז יהיה עלינו לשאול עוד שאלה כדי להגיע לעלים מונוכרומטיים (טהורים).

העץ האדום הוא overfitting, מפני שהוא מתאים מידי ל-training data שלנו, לכן יש מוטיבציה לבצע pruning באמצעות validation set (למשל).

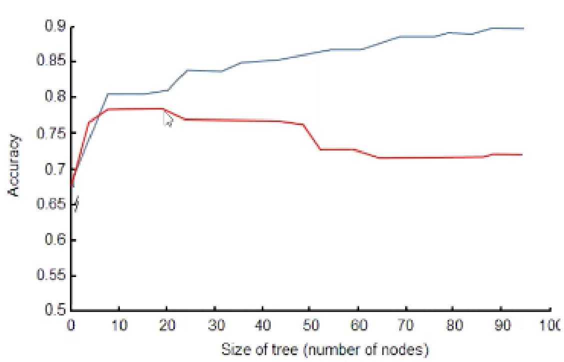


:Post-Pruning using a Validation Set

- נפצל את הדאטה לקבוצת אימון training set ולקבוצת בדיקה test set.
- נבנה עץ על ה-training data (עם או בלי בסיס לקיצוץ, למשל chi-square)
- נקצץ צמתים אשר עבורם הביצועים של הסיווג/הקלסיפיקציה טובים יותר על הטסט סט כאשר מבטלים את הילדים שלהם.



הגרף הבא מתאר את **הדיוק בקלסיפיקציה כפונקציה של גודל העץ** (מבחינת מספר הצמתים) כפי שניתן לראות הקו הכחול מתאר את הדיוק עבור ה-training data שהולך ומשתפר ככל שמגדילים את העץ ובכך נוצר overfitting. הגרף האדום מתאר את הדיוק של המודל שלנו עבור ה-validation set. נשים לב שבשלב מסויים, ככל שנעשית התאמה עבור ה-training set, הדיוק עבור ה-test set הולך ופוחת, מה שמתאים לנו להגדרה של overfitting.



	Less than 1.70m	1.70-1.90	Taller than 1.90
Women	4/6	4/6	8/9
Men	1/2	1/2	23/27

MAP classification (next week)

Classify an instance with observed properties $\vec{x}$ as

$$\underset{i}{\text{argmax}}\, P(\vec{x}|A_i)P(A_i)$$

הכנה להרצאה הבאה – A prolog on Bayesian Learning

פרדוקס סימפסון: נאספו נתונים אודות שחקני כדורסל שהצליחו לקלוע 5/5 סלים מהקו.

הדאטה מוצגת בטבלה משמאל ומפולחת על פי גובה ומין.

האינטואיציה אומרת – פלח גדול יותר של הנשים בכל קטגוריה צלחו את הקליעה, ולכן הנשים יותר טובות לפי אינטואיציה זו.

אבל כאשר סוכמים את מס׳ הנשים שצלחו לעומת הגברים שצלחו, נקבל **שיותר גברים צלחו.**

המשך הכנה להרצאה הבאה – Parameter and model estimation

- **Density Estimation** – בהינתן נתונים נרצה להעריך את פונקציית הצפיפות של הדאטה = PDF Probability Density Function.

הערכת סבירות מקסימלית – MLE = Maximum Likelihood Estimation

- גישה ישירה עבור הערכת פרמטרים אשר עובדת באופן ישיר על מקרים פשוטים ויוצרת בסיס רעיוני עבור רוב הגישות העוסקות בהערכת פרמטרים.
- בהינתן סט דגימות D = {x1,…xm} ומודל ווקטור מועמד של הפרמטרים של המודל הזה, טטה.
- נגדיר את הסבירות (Likelihood) של כל מודל מועמד בהינתן הדאטה להיות L(theta | D), למען זאת נשתמש ב- P(D | theta) ההסתברות לראות את הדאטה בהינתן הוקטור theta.
- ה-log-likelihood של המודל בהינתן הדאטה: LL(theta) = log P(D | theta)
- נרצה למקסם את הסבירות, אם וקטור טטה ממקסם סבירות הוא גם ימקסם את לוג-הסבירות. ולכן ב-MLE אנחנו מחפשים את:

$$\theta^* = \underset{\theta\in\Omega}{\text{argmax}}\, LL(\theta)$$

MLE for independent identically distributed instances

- לרוב נניח כי דגימות הדאטה נוצרות ממשתנים מקריים שהם independent identically distributed (i.i.d).
- לכן, עלינו למצוא את:

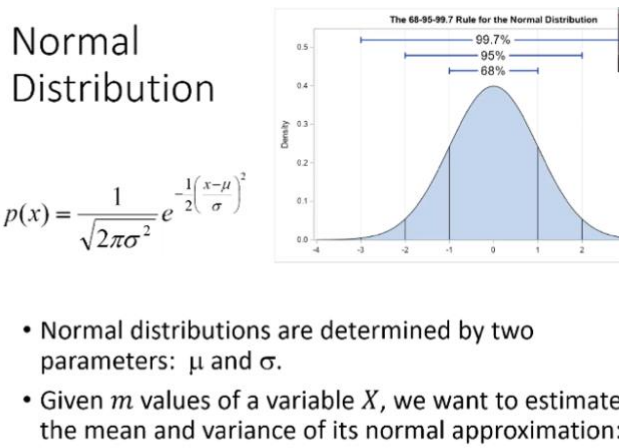
$$\theta^* = \underset{\theta\in\Omega}{\text{argmax}} \sum_{i=1}^{m} P(x_i|\theta)$$

דוגמה של הטלת מטבע (לא נדע מהו ה-p של המטבע וננסה להעריך אותו)

- נניח כי יש בידינו מטבע בעל הסתברות p להיות heads (H) והסתברות q=1-p להיות tails (T).
- Observation: נזרוק את המטבע m פעמים, ונתבונן בקבוצה של H-ים ו-T-ים.
- $$L(\Theta) = \log P(D\,|\,\Theta) = \log p^m (1-p)^{N-m}$$
$$= m\log p + (N-m)\log(1-p)$$
- (חסר NchooseM = טעות, אבל זהו קבוע, זה לא ישנה את הטטה הממקסמת ולכן נזניח אותו)
- נרצה למצוא את ה-p שמביא את הנוסחה לעיל למקסימום ולכן גוזרים את הלייקליהוד לפי p ומשווים ל-0:
- $$\frac{dL(\Theta)}{dp} = \frac{d(m\log p + (N-m)\log(1-p))}{dp} = \frac{m}{p} - \frac{N-m}{1-p} = 0$$
- $\Rightarrow\ p = m/N$ זהו ה-p אשר מביא למקסימום.

דוגמה של ההתפלגות הגאוסיאנית Normal Distribution

נתבונן במשפחת המשתנים שמתפלגים נורמלית/גאוסיאני אשר מאופיינים על ידי שני פרמטרים מיו = תוחלת, וסיגמא = סטיית תקן. (ככל שהסיגמא יותר קטנה = פעמון יותר צר). נתבונן במדידות D = {x1,…xn} אשר אנו מניחים שהן מגיעות מתוך ההתפלגות הנורמלית. במקרה זה מתקיים כי $\theta = (\mu, \sigma^2)$. וכן, פונקציית ה-likelihood: $L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$.



יותר נוח לעבוד עם ה-log-likelihood:

$$LL(\theta) = -n\ln\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

כדי למצוא את המקסימום נחשב את הגרדיאנט ונשווה ל-0:

To find a max point for this function we set the gradient to 0:

$$\frac{\partial}{\partial\mu} \ :\ \sum_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0 \ \Rightarrow\ \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial(\sigma^2)} \ :\ -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^4} = 0 \ \Rightarrow\ \hat{\sigma} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)^{1/2}$$

עבור מיו-כובע נוכל לחשב מפני שהדגימות נתונות לנו. לכן נקבל הערכה למיו.
עבור סיגמה-האט נציב את המיו-כובע בתוך הנוסחה שהתקבלה.
לכן בהכרח קודם חייב לחשב את ההערכה לתוחלת ונציב אותה בהערכה של סטיית התקן (סיגמה).

דוגמה נוספת – התפרצות הר הגעש Old Faithful Wyoming

נלמד אלגוריתם EM על מנת להעריך את ההתפרצויות הללו.