

Concepte generale. Clasificatorul Bayes Naiv. Măsurarea performanței.

Prof. Dr. Radu Ionescu
raducu.ionescu@gmail.com

Facultatea de Matematică și Informatică
Universitatea din București

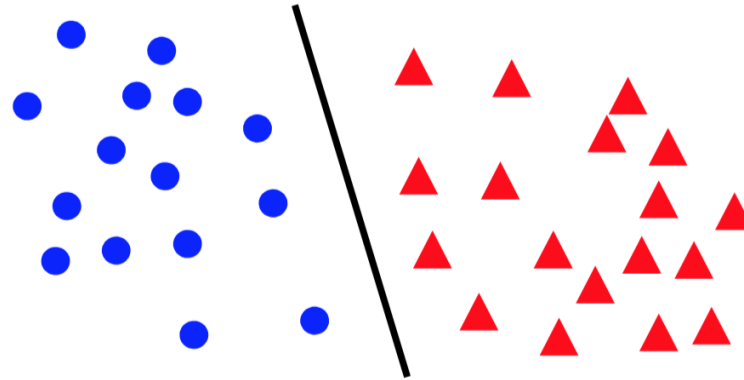
Paradigme ale învățării

- Învățare supervizată (supervised learning)
- Învățare nesupervizată (unsupervised learning)
- Învățare semi-supervizată (semi-supervised learning)
- Învățare ranforsată (reinforcement learning)

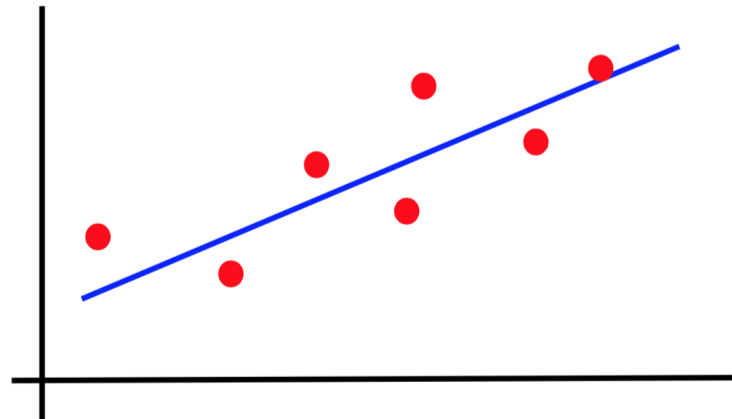
- Paradigme non-standard:
 - Învățarea activă (active learning)
 - Învățare prin transfer (transfer learning)

Formele canonice ale problemelor de învățare supervizată

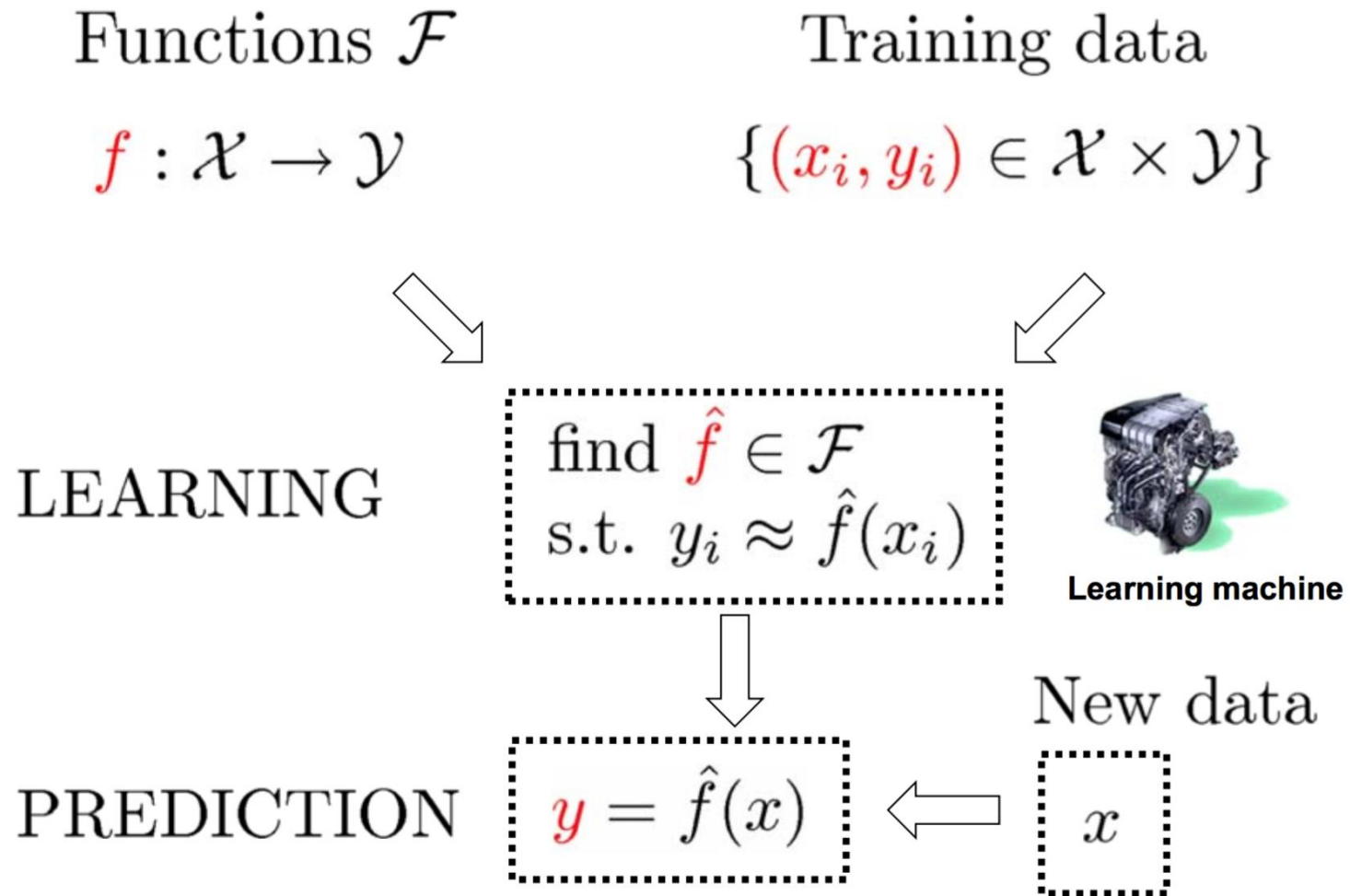
- Clasificare



- Regresie



Paradigma de învățare supervizată



Pașii necesari pentru învățare supervizată

- **Definirea** problemei de învățare supervizată

- **Colectarea datelor**

Pornim cu datele de antrenare, pentru care știm etichetele corecte (de la un professor sau oracol)

- **Reprezentarea datelor**

Alegem cum să reprezentăm datele

- **Modelarea**

Alegerea spațiului de ipoteze: $H = \{g: X \rightarrow Y\}$

- **Învățarea / Estimarea parametrilor**

Găsirea celei mai bune ipoteze din spațiul ales

- **Selectarea modelului**

Încercăm mai multe modele și îl păstrăm pe cel mai bun

- Dacă rezultatele sunt mulțumitoare atunci ne oprim

Altfel rafinăm unul sau mai mulți pași anteriori

Clasificare între Banana și Furbish

- **Date de antrenare**
- Banana language:
 - baboi, bananonina, bello, hana, stupa
- Furbish:
 - doo, dah, toh, yoo, dah-boo, ee-tay
- Date de test: gelato
- Care este limba?
- De ce?
- **Învățarea este grea fără a stabili un spațiu de ipoteze H!**



Antrenare versus testare

- Ce ne dorim?
 - Performanță bună (pierdere scăzută) pe datele de antrenare?
 - Nu, performanță bună pe datele de test (nevăzute)
- Date de antrenare:
 - $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$
 - Sunt date pentru a învăța funcția de mapare f
- Date de testare:
 - $\{x_1, x_2, \dots, x_M\}$
 - Folosite pentru a vedea cât de bine am învățat

Funcția de eroare / de pierdere

- Cum măsurăm performanța?
- Regresie:
 - Media pătratelor erorilor
 - Media erorilor în valoare absolută
- Clasificare:
 - Numărul de clasificări greșite (misclassification error)
 - Pentru clasificare binară:
True Positive, False Positive, True Negative, False Negative
 - Pentru clasificare în mai multe clase:
Matricea de confuzie

Erori

- Eroarea de generalizare (generalization error):

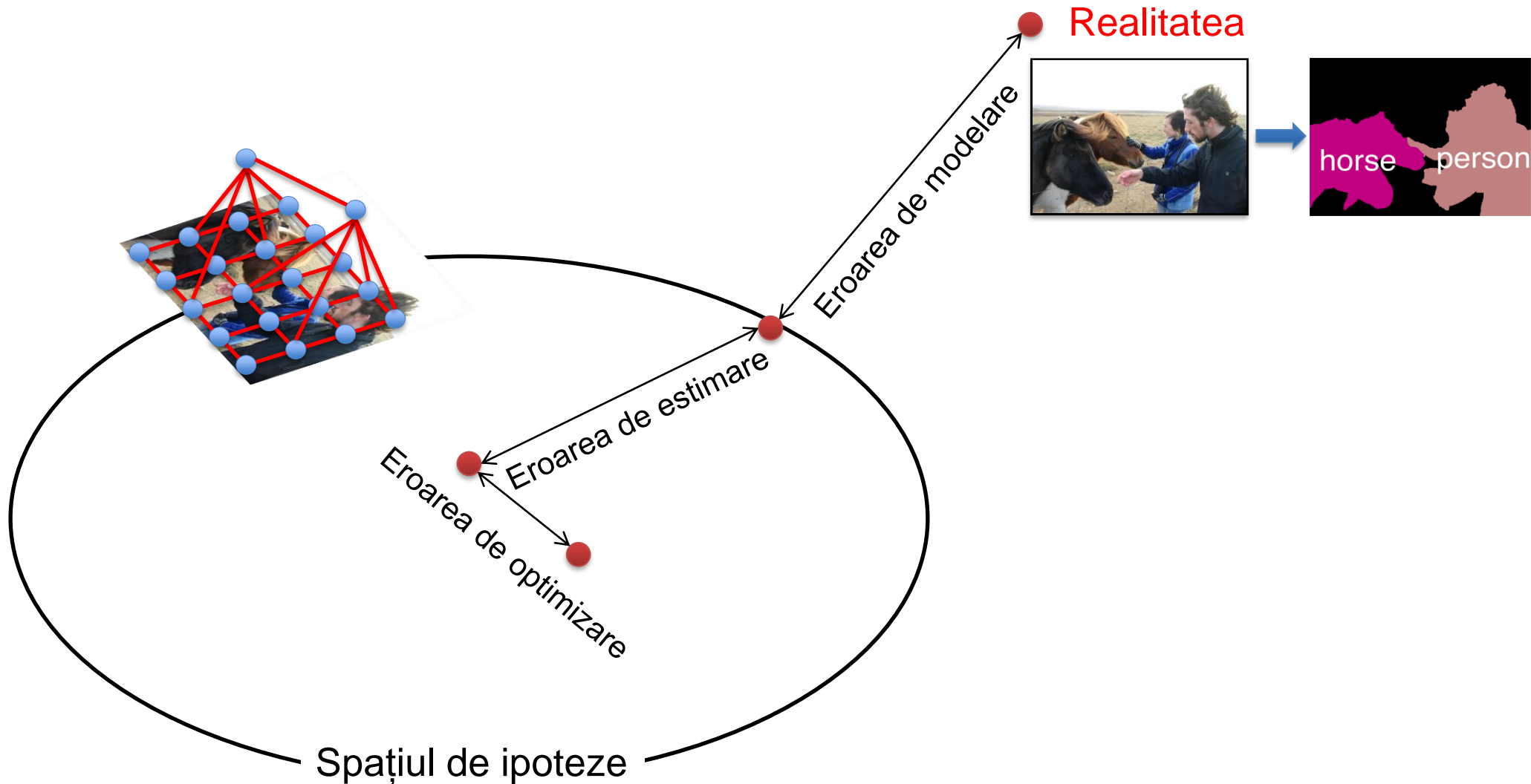
$$\mathcal{E}(h) = \int_{X \times Y} V(h(x), y) \rho(x, y) dx dy$$

- Probabilitatea comună $\rho(x, y)$ este de obicei necunoscută
- Atunci calculăm eroare empirică (empirical error):

$$E(h) = \frac{1}{n} \sum_{i=1}^n V(h(x_i), y_i)$$

- Estimăm eroarea empirică pe datele de antrenare sau pe cele de test?
- **Nu este corect să raportăm eroarea pe datele de antrenare!**

Descompunerea erorii

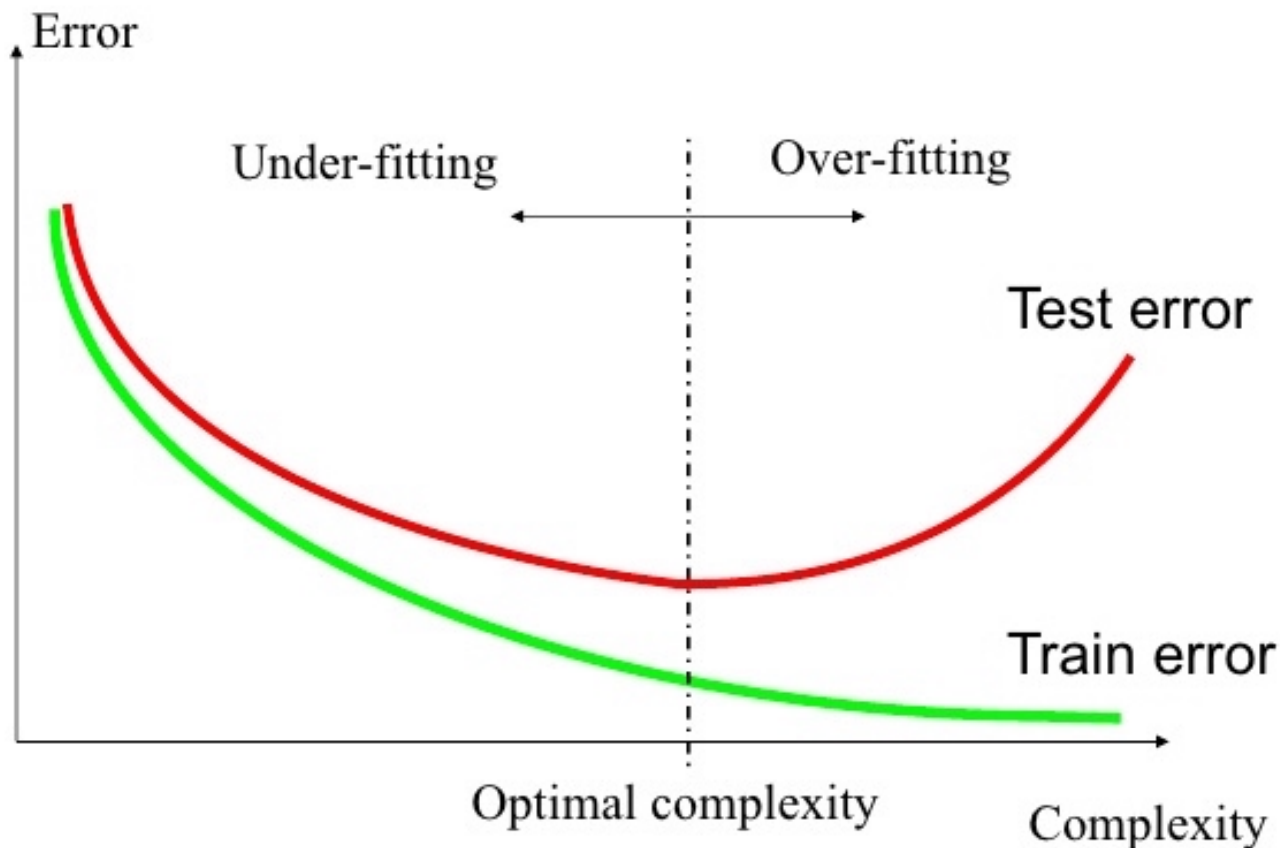


Descompunerea erorii

- Eroare de modelare
 - Am încercat să modelăm realitatea cu un spațiu de ipoteze
- Eroarea de estimare
 - Am încercat să antrenăm un model cu o mulțime finite de date
- Eroarea de optimizare
 - Nu am reușit să optimizăm funcția până în punctul optim

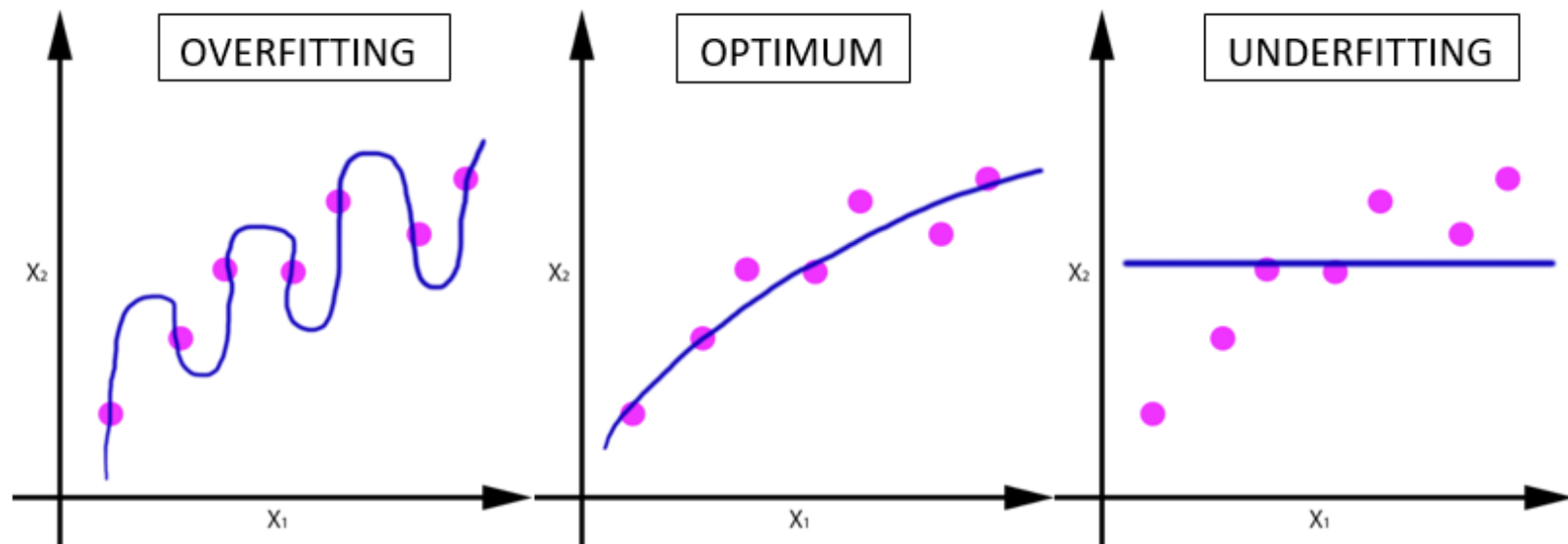
Underfitting versus overfitting

- Problema cea mai importantă a învățării?
- Îmbunătățirea capacității de generalizare



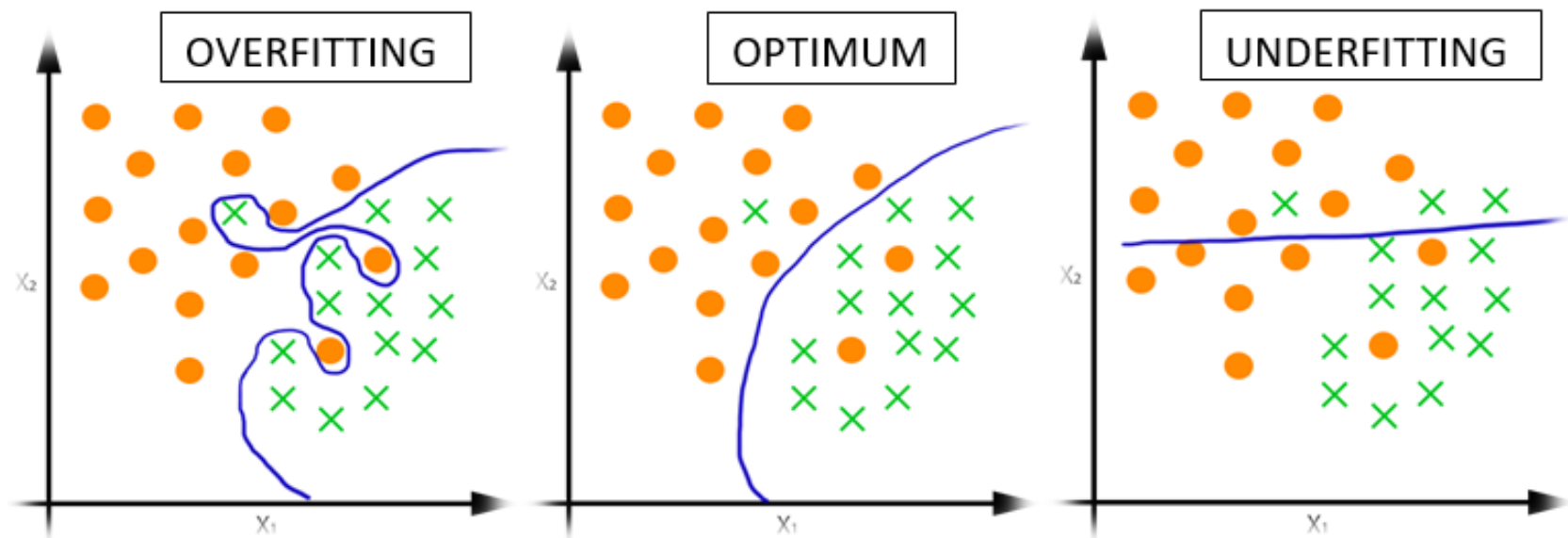
Underfitting versus overfitting

- Exemplu 1: problemă de regresie



Underfitting versus overfitting

- Exemplu 2: problemă de clasificare



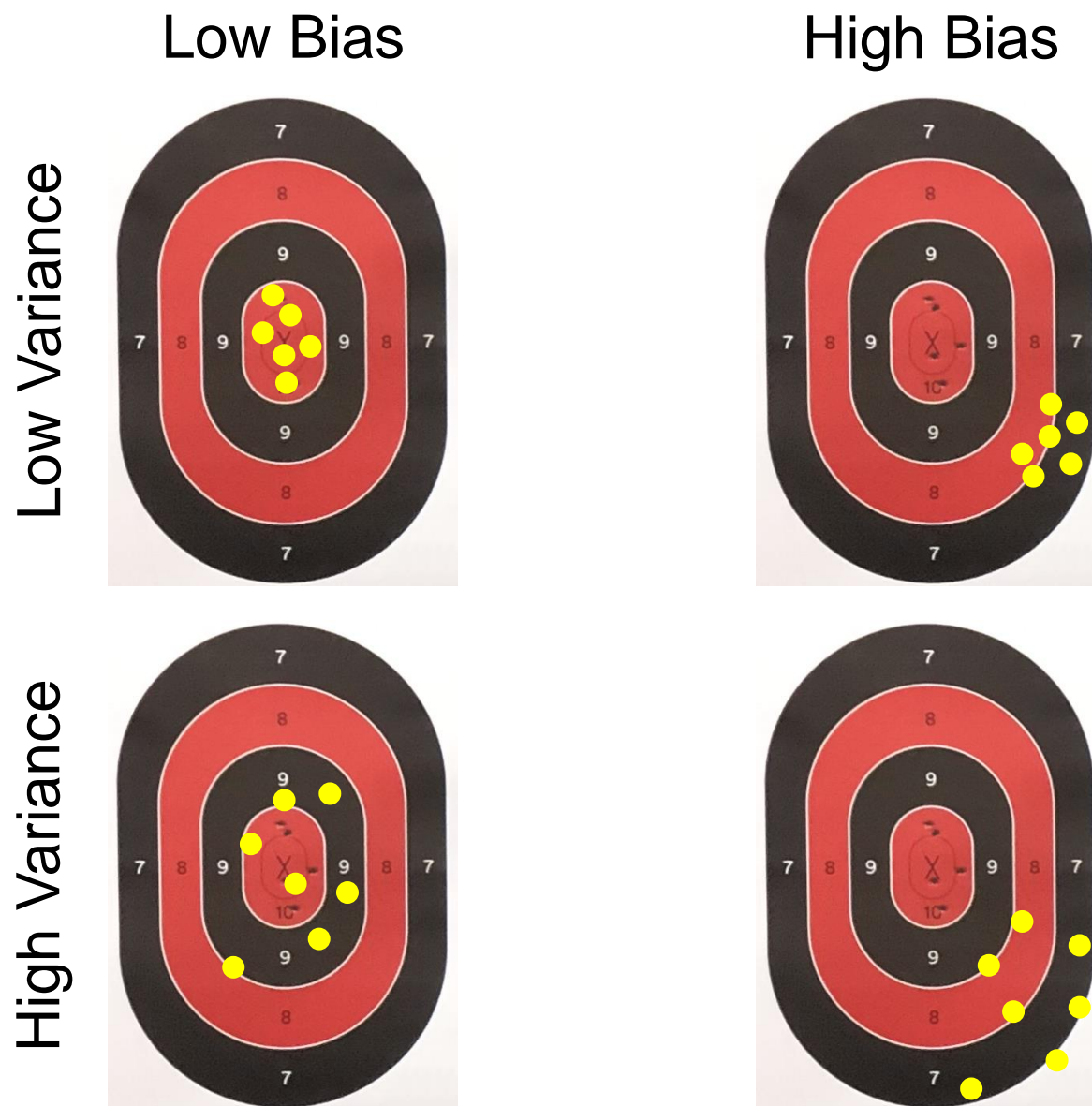
Bias-Variance Trade-off

- Bias
 - Eroare sistematică care provine din inabilitatea modelului de a învăța adevărata relație dintre trăsături și etichete (underfitting)
 - Poate fi corectată prin creșterea complexității modelului
- Variance
 - Eroare aleatoare care provine din sensibilitatea ridicată la mici fluctuații din date, cauzată de faptul că modelul a învățat și zgomotul din datele de antrenare (overfitting)
 - Poate fi corectată prin adăugarea de exemple de antrenare sau prin scăderea complexității modelului

Bias-Variance Trade-off



Bias-Variance Trade-off



Abordarea procedurală

- Etapa de antrenare:
 - Date neprelucrate $\rightarrow x$
(extragerea trăsăturilor / caracteristicilor = feature extraction)
 - Date de antrenare $\{(x,y)\} \rightarrow f$
(învăţare)
- Etapa de testare:
 - Date neprelucrate $\rightarrow x$
(extragerea trăsăturilor)
 - Date de testare $x \rightarrow f(x)$
(aplicarea funcţiei, calcularea erorii)

Abordarea statistică

- Folosim probabilități:
 - x și y sunt variabile aleatoare
 - $D = (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N) \sim P(X, Y)$
- Presupunem că datele sunt i.i.d. (independent și identic distribuite):
 - Datele de antrenare și testare sunt generate i.i.d. din $P(X, Y)$
 - Învățăm pe setul de antrenare
 - Sperăm ca modelul să **generalizeze** pe datele de test

Concepte

- Capacitatea modelului
 - Cât de larg este spațiul de ipoteze H ?
 - Este sau nu restrâns spațiul de funcții?
- Supra-învățare (overfitting)
 - f funcționează bine pe datele de antrenare
 - Dar foarte slab pe datele de testare
- Capacitatea de generalizare
 - Abilitatea de a obține eroare mică pe datele noi de test

Garanții

- Simplificând 20 de ani de cercetare din Teoria Învățării...
- Dacă:
 - Avem suficiente date de antrenare D
 - Și spațiul de ipoteze H nu este foarte complex
- atunci **probabil** că modelul va avea capacitate de generalizare

Probabilități (recapitulare)

- A este un eveniment nedeterminist:
A = “Simona Halep va câștiga Roland Garros”

- Ce înseamnă $P(A)$?

- Abordarea statistică:

$$\lim_{N \rightarrow \infty} \frac{\#(A = \text{true})}{N}$$

- Frecvența la limită a unui eveniment repetabil și nedeterminist

- Abordarea Bayesiană:

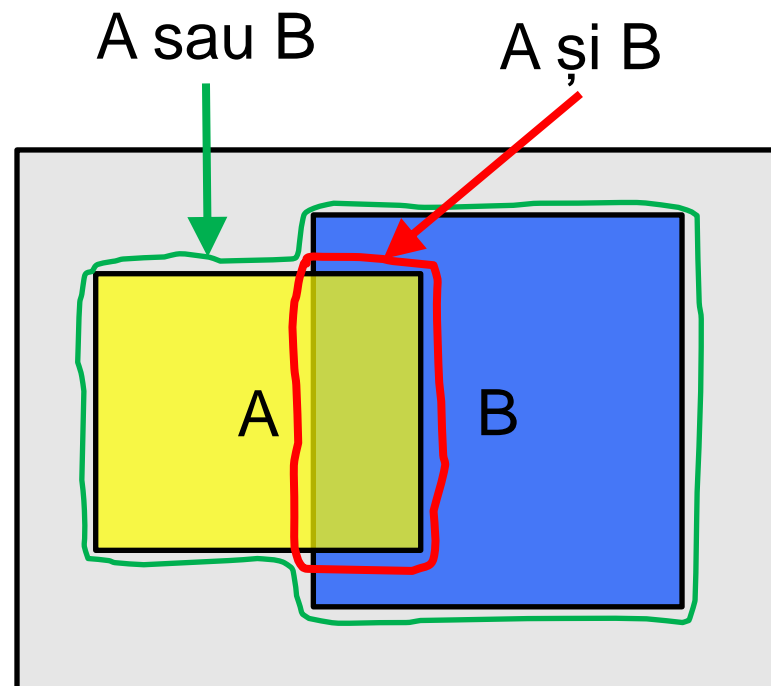
- $P(A)$ este ceea ce “credem” despre A

- Abordarea economică:

- $P(A)$ ne spune cât de mult “pariem” dacă alegem A

Axiomele Probabilității (recapitulare)

- $0 \leq P(A) \leq 1$
- $P(\emptyset) = 0$
- $P(\Omega) = 1$
- $P(A \text{ sau } B) = P(A) + P(B) - P(A \text{ și } B)$



Probabilități condiționate (recapitulare)

$$P(Y = y \mid X = x)$$

- Ce să credem despre $Y = y$, dacă știm că $X = x$?
- $P(\text{Simona Halep va câștiga Roland Garros 2020})?$
- Dacă știm următoarele:
 - În 2018, Simona Halep a câștigat Roland Garros
 - Simona Halep a pierdut două finale de Roland Garros
 - Simona se află pe poziția a doua în clasamentul WTA
 - În 2018, Naomi Osaka (poziția întâi WTA) a pierdut în turul al treilea

Probabilități condiționate (recapitulare)

- $P(A | B)$ = În cazurile în care B este adevărat, proporția în care A este adevărat

- Exemplu:

➤ D: “Am dureri de cap”

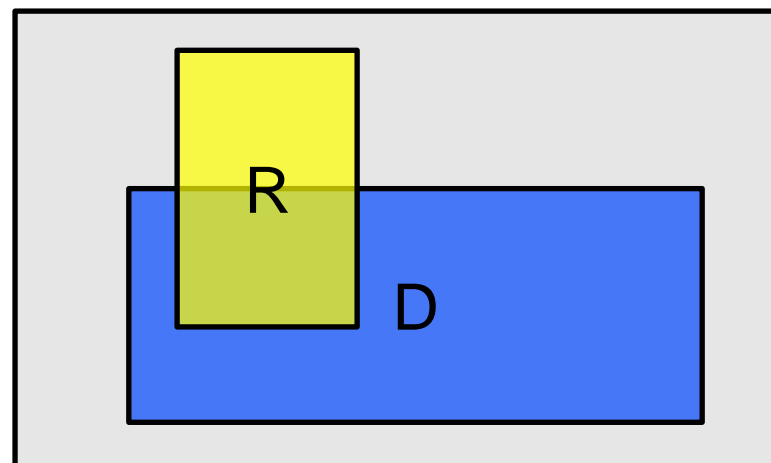
➤ R: “Sunt răcit”

- $P(D) = \frac{1}{10}$

- $P(R) = \frac{1}{40}$

- $P(D | R) = \frac{1}{2}$

- Durerile de cap sunt rare și răceala este și mai rară, dar dacă ești răcit atunci sunt 50% șanse să ai dureri de cap

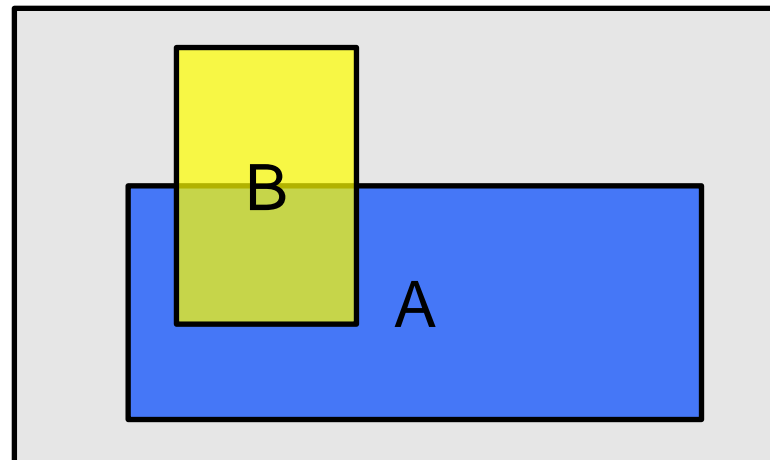


Regula Bayes

$$P(B | A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A | B) P(B)}{P(A)}$$



- Thomas Bayes "An Essay towards solving a Problem in the Doctrine of Chances" Royal Society, 1763.
- Simplu de înțeles dacă vă gândiți la arii



Regula Bayes

Concepte:

- Probabilitate

- Cât de bine explică datele o anumită ipoteză?

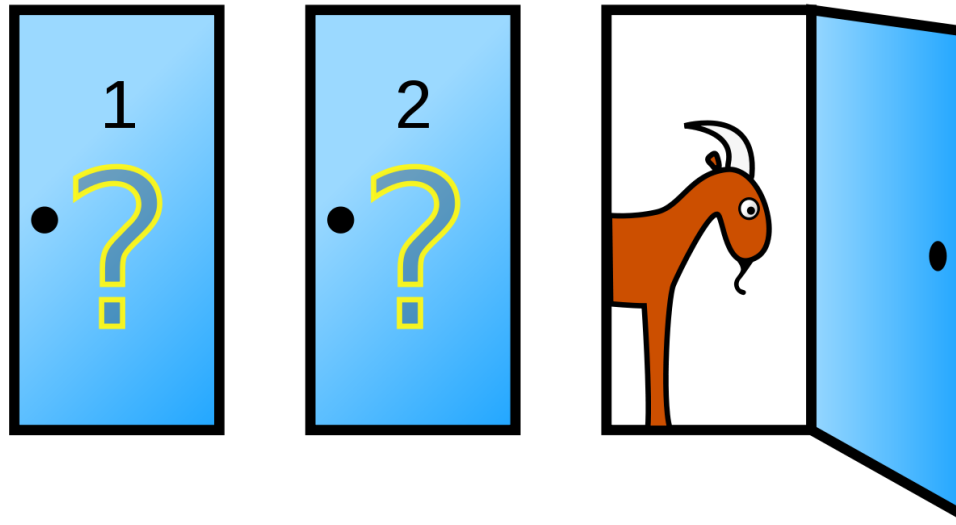
- Informații apriori

- Ce credem înainte de a vedea datele?

- Informații aposteriori

- Ce credem după ce vedem datele?

Problema Monty Hall



- Sunt 3 uși numerotate cu 1, 2, 3.
 - Un premiu mare (o mașină) este ascunsă în spatele unei uși. Celelalte două uși au câte o capră.
 - Trebuie să alegem o ușă.
 - Să presupunem că alegem poarta 1. Gazda deschide poarta 3, arătând capra din spate. Ce alegem mai departe?
- (a) Rămânem cu alegerea inițială (poarta 1);
- (b) Schimbăm și alegem poarta 2;
- (c) Este vreo diferență?

Problema Monty Hall

- $H = i$ denotă ipoteza “premiul este după ușa i ”. Apriori toate cele 3 uși sunt egal probabile să ascundă premiul:

$$P(H = 1) = P(H = 2) = P(H = 3) = \frac{1}{3}$$

- Alegem poarta 1.
- Dacă premiul este în spatele ușii 1, gazda este indiferentă și va alege ușile 2 sau 3 cu probabilitate egală:

$$P(U = 2 \mid H = 1) = \frac{1}{2}, P(U = 3 \mid H = 1) = \frac{1}{2}$$

- Dacă premiul este în spatele ușii 2 (respectiv 3), gazda alege ușa 3 (respectiv 2):

$$P(U = 2 \mid H = 2) = 0, P(U = 3 \mid H = 2) = 1$$

$$P(U = 2 \mid H = 3) = 1, P(U = 3 \mid H = 3) = 0$$

- Gazda deschide poarta 3 ($U=3$), descoperind capra. Observația este $U=3$. Premiul este în spatele ușii 1 sau 2?

Problema Monty Hall

$$\begin{aligned}P(H = 1) &= P(H = 2) = P(H = 3) = \frac{1}{3} \\P(U = 2 \mid H = 1) &= \frac{1}{2}, P(U = 3 \mid H = 1) = \frac{1}{2} \\P(U = 2 \mid H = 2) &= 0, P(U = 3 \mid H = 2) = 1 \\P(U = 2 \mid H = 3) &= 1, P(U = 3 \mid H = 3) = 0\end{aligned}$$

- Aplicăm regula Bayes:

$$P(H = 1 \mid U = 3) = \frac{P(U = 3 \mid H = 1) P(H = 1)}{P(U = 3)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

$$P(H = 2 \mid U = 3) = \frac{P(U = 3 \mid H = 2) P(H = 2)}{P(U = 3)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Clasificatorul optimal

- **Învățăm:** $h: \mathbf{X} \rightarrow Y$
 - \mathbf{X} – trăsături
 - Y – etichete
- Presupunând cunoscută $P(Y|\mathbf{X})$, cum clasificăm datele?
 - Aplicăm clasificatorul Bayes:

$$y^* = h^*(x) = \operatorname{argmax}_y P(Y = y \mid X = x)$$

- **De ce?**

Clasificatorul optimal

- **Teoremă:** Clasificatorul Bayes h_{Bayes} este optim!

- Adică:

$$error_{true}(h_{\text{Bayes}}) \leq error_{true}(h), \forall h$$

- **Eroarea Bayes** este cea mai mica eroare posibilă:

$$error_{\text{Bayes}} = 1 - \sum_{y \neq y^*} \int_{x \in H_i} P(y | x) P(x) dx$$

Clasificatorul optimal

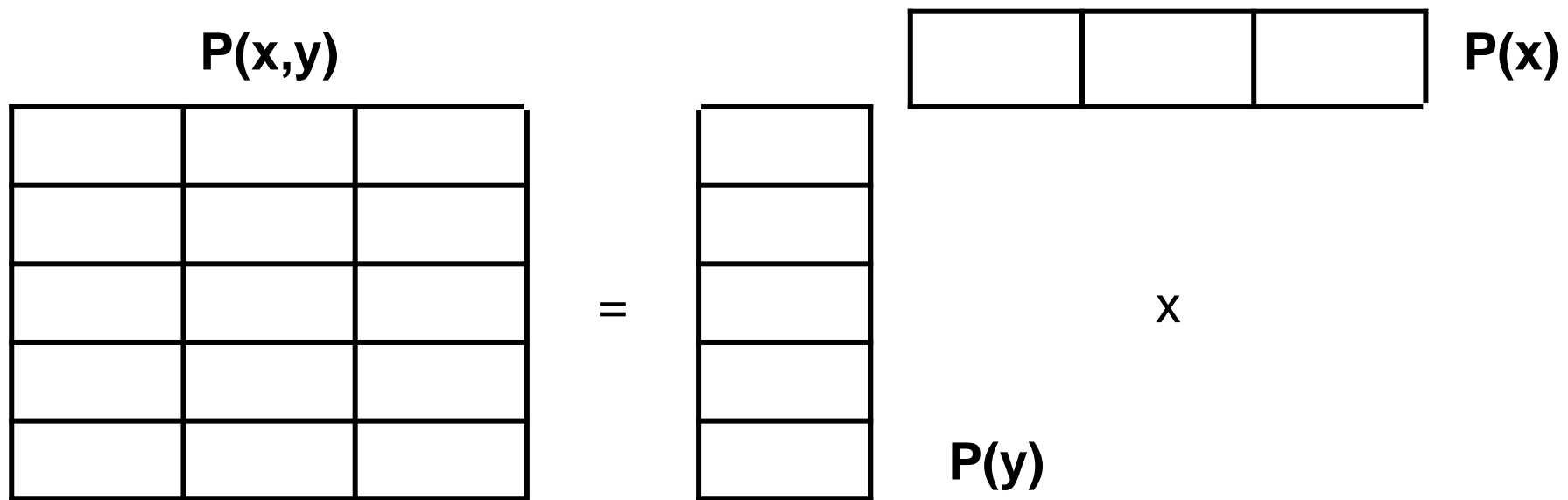
- Cât de greu este să învățăm clasificadorul optimal?
 - Dar pentru date categorice?
- Cum reprezentăm datele? Câți parametrii trebuie estimați?
 - Probabilitatea apriori a claselor $P(Y)$:
Presupunem că Y este compus din k clase
 - Probabilitatea $P(\mathbf{X} | Y)$:
Presupunem că \mathbf{X} este compus din n trăsături binare

Model complex → Avem varianță mare cu date limitate!

Soluție: considerăm că trăsăturile sunt independente

- Două variabile sunt independente dacă și numai dacă:

$$P(x, y) = P(x) P(y)$$



- Două variabile sunt independente condiționat dacă, fiind dată o a treia variabilă, avem:

$$P(x, y | z) = P(x | z) P(y | z)$$

Clasificatorul Naïve Bayes

- Presupunerea Naïve Bayes:

- Trăsăturile sunt independente:

$$P(X_1, X_2 | Y) = P(X_1 | Y)P(X_2 | Y)$$

- Mai general:

$$P(X_1 \dots X_n | Y) = \prod_i P(X_i | Y)$$

- Câți parametrii trebuie estimați acum?

- Presupunem că **X** este compus din n trăsături binare

- Redus de la 2^n la $2 \cdot n$

Clasificatorul Naïve Bayes

- Fiind date:
 - Probabilitatea apriori a claselor $P(Y)$
 - n trăsături independente **X** condiționate de Y
 - Pentru fiecare X_i , probabilitatea $P(X_i | Y)$

- Regula de decizie Naïve Bayes este:

$$h_{NB}(x) = \underset{y}{\operatorname{argmax}} P(y) P(x_1, \dots, x_n | y)$$

$$h_{NB}(x) = \underset{y}{\operatorname{argmax}} P(y) \prod_i P(x_i | y)$$

- În practică folosim sumă de log!
- Dacă presupunerea este adevărată, NB este clasificadorul optimal!

Estimarea parametrilor NB

- Se aplică metoda aproximării verosimilității maxime (Maximum Likelihood Estimation)
 - Fiind dat setul de antrenare, calculăm numărul de exemple pentru care $A=a$ și $B=b$:

`count(A=a, B=b)`

- Estimarea parametrilor:
 - Probabilitatea apriori a fiecărei clase: $P(Y = y) = \dots$
 - Probabilitatea condiționată de clase: $P(X_i = x_i | Y = y) = \dots$

Încălcarea presupunerii NB

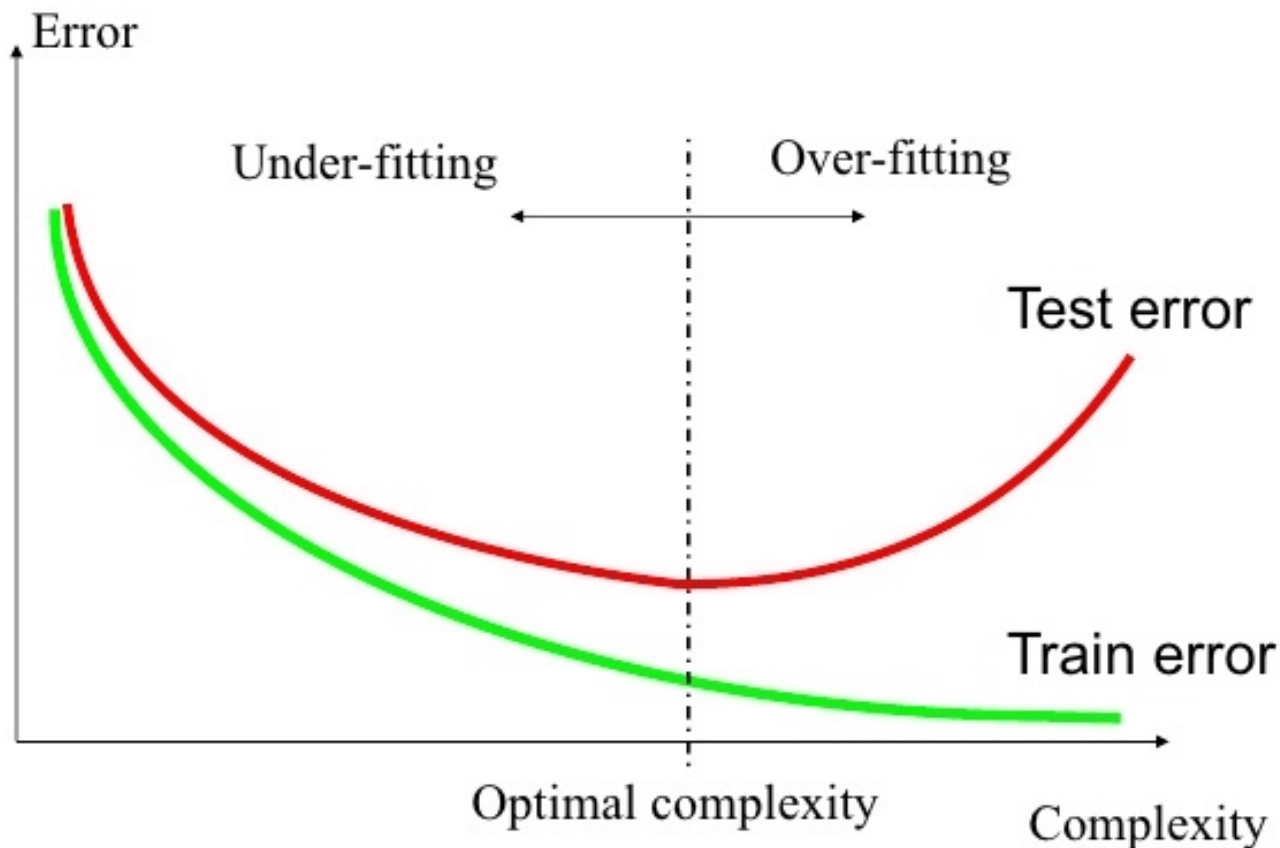
- De obicei, trăsăturile nu sunt independente condiționat:

$$P(X_1 \dots X_n | Y) \neq \prod_i P(X_i | Y)$$

- Probabilitățile $P(Y|\mathbf{X})$ sunt deseori 0 sau 1
- Totuși, clasificatorul NB este foarte popular
 - Deorece se descurcă bine, chiar dacă presupunerea este încălcată

Underfitting versus overfitting

- Îmbunătățirea capacității de generalizare



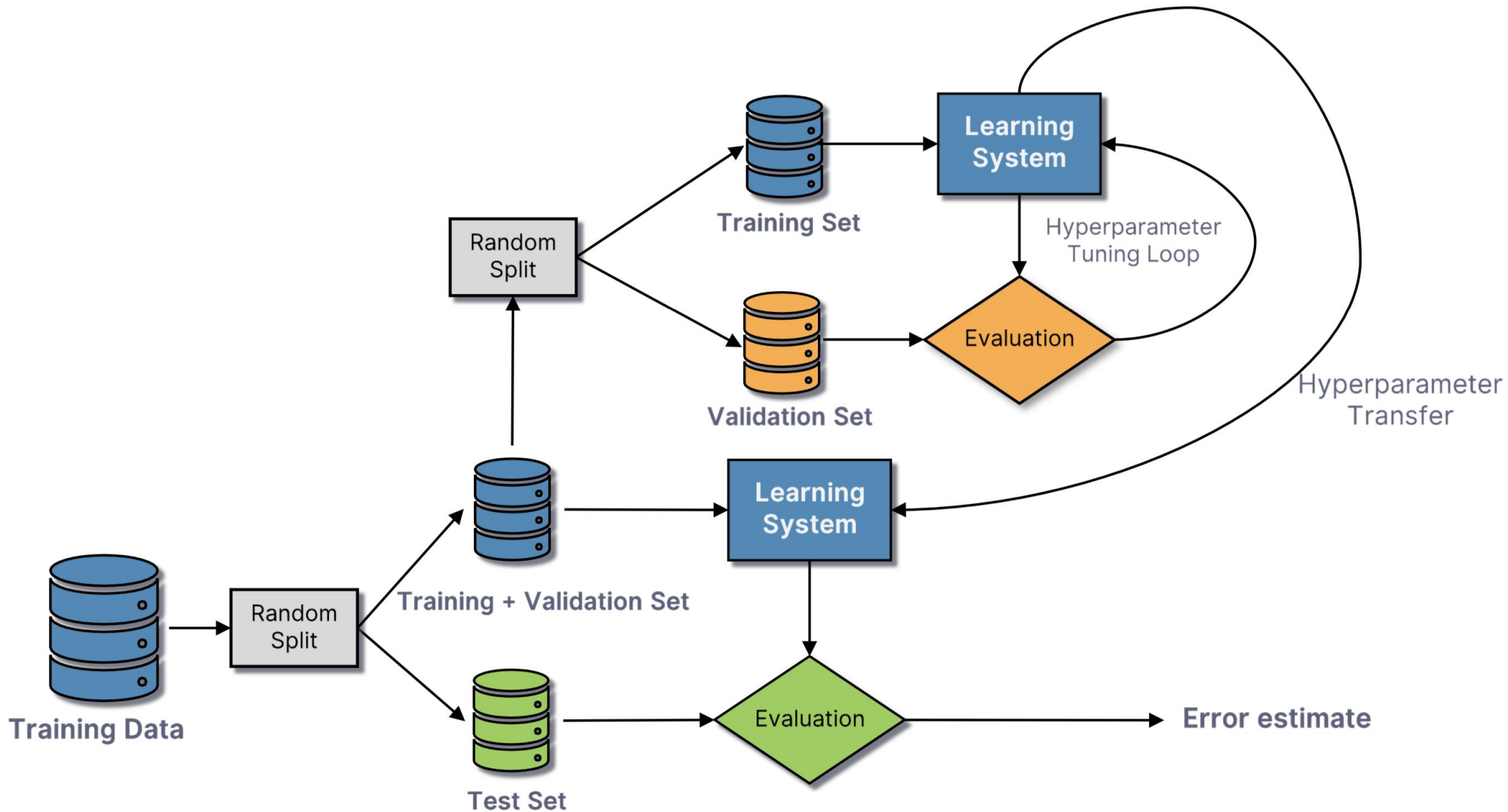
Împărțirea datelor în date de antrenare, validare și test

- Pentru a construi un model cât mai performant, trebuie să îl testăm pe date “necunoscute”
 - O posibilă abordare (atunci când avem la dispoziție multe date):
 - 50% exemple pentru antrenare
 - 25% exemple pentru validare
 - 25% exemple pentru testare
- (procentele pot să varieze)

De ce nu este suficient să împărțim datele în train și test?

- Utilizarea repetată a unei împărțiri atunci când încercăm diverși hiperparametrii poate să “uzeze” setul de de test:
 - **Facem overfitting în spațiul hiperparametrilor!**
- Obținem o estimare mai bună a erorii dacă tunăm hiperparametrii pe un set diferit, anume setul de validare

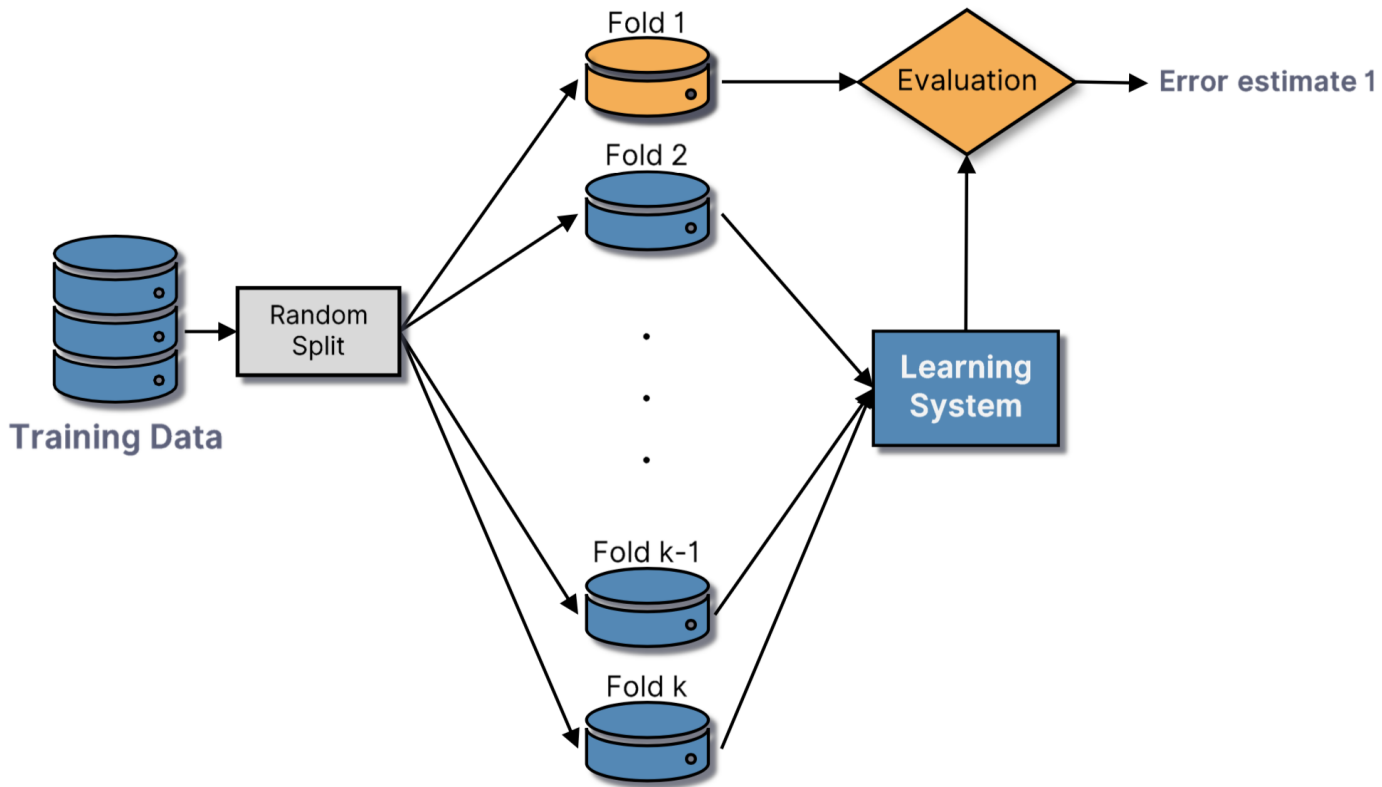
Training, validation, test



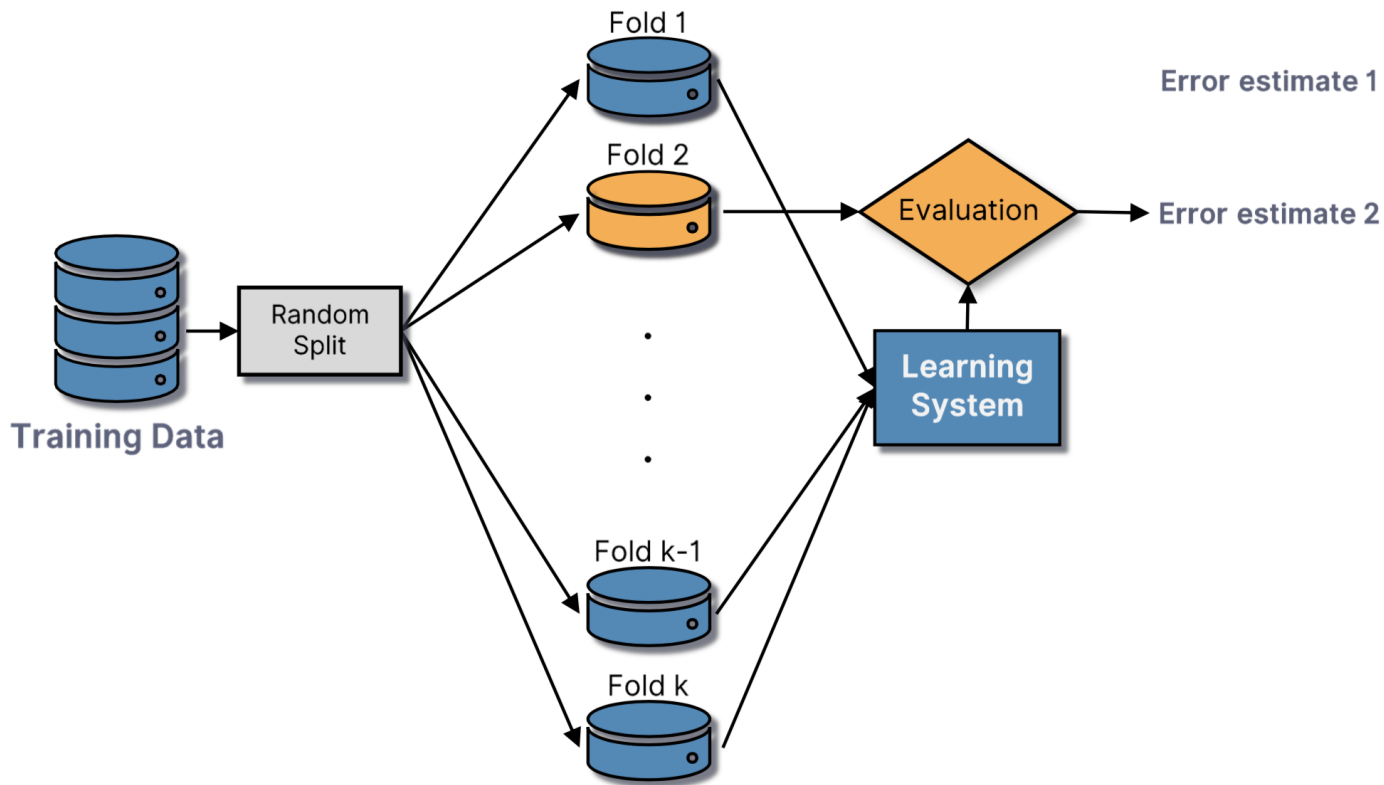
Cross-validation

- O altă abordare (funcționează bine cu data puține):
 - Împărțim datelor în k părțile egale (fold-uri)
 - Antrenăm pe $k-1$ fold-uri și testăm pe fold-ul dat deoparte
 - Repetăm de k ori
 - Calculăm media rezultatelor
- Atunci când numărul de fold-uri este egal cu numărul de exemple:
 - Leave-one-out cross-validation

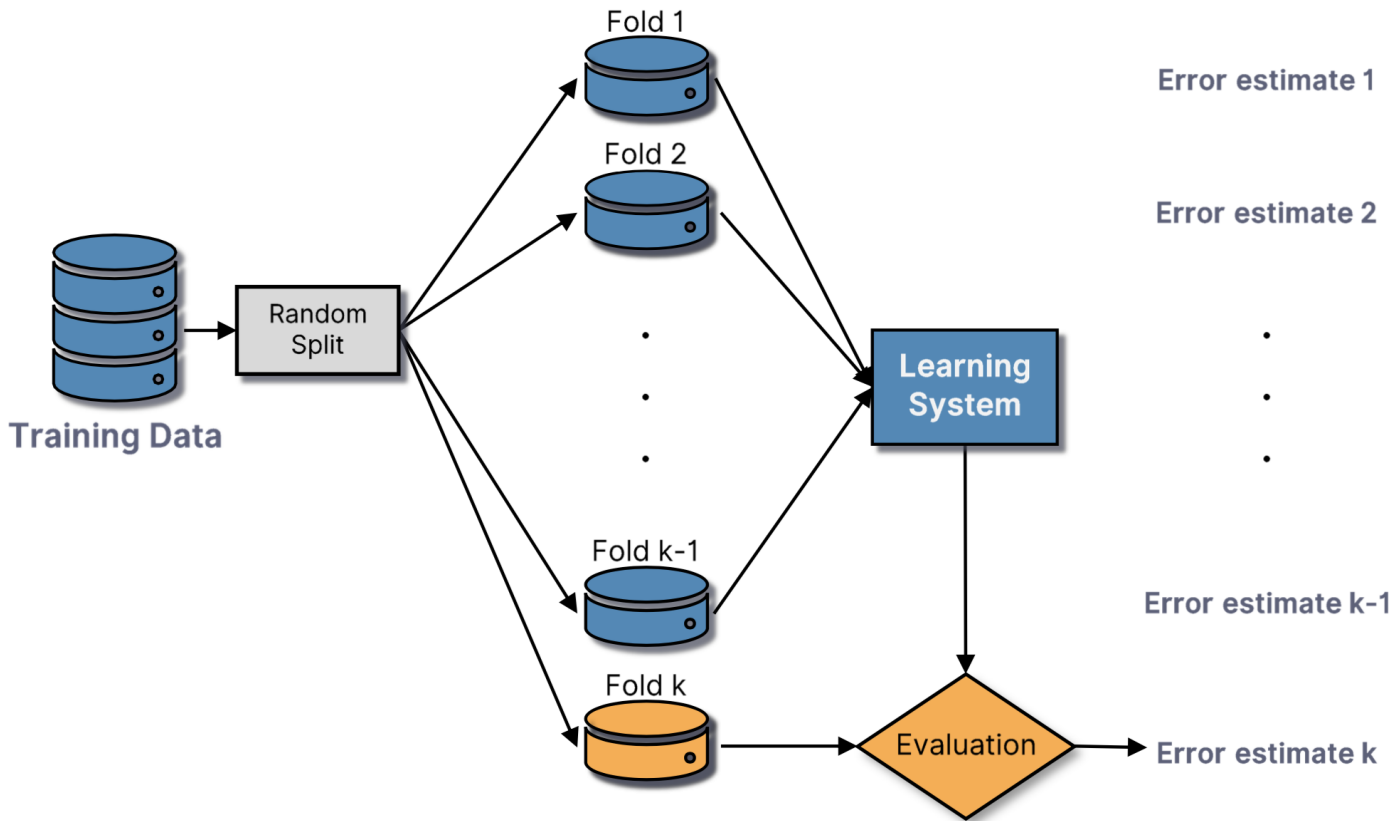
Cross-validation



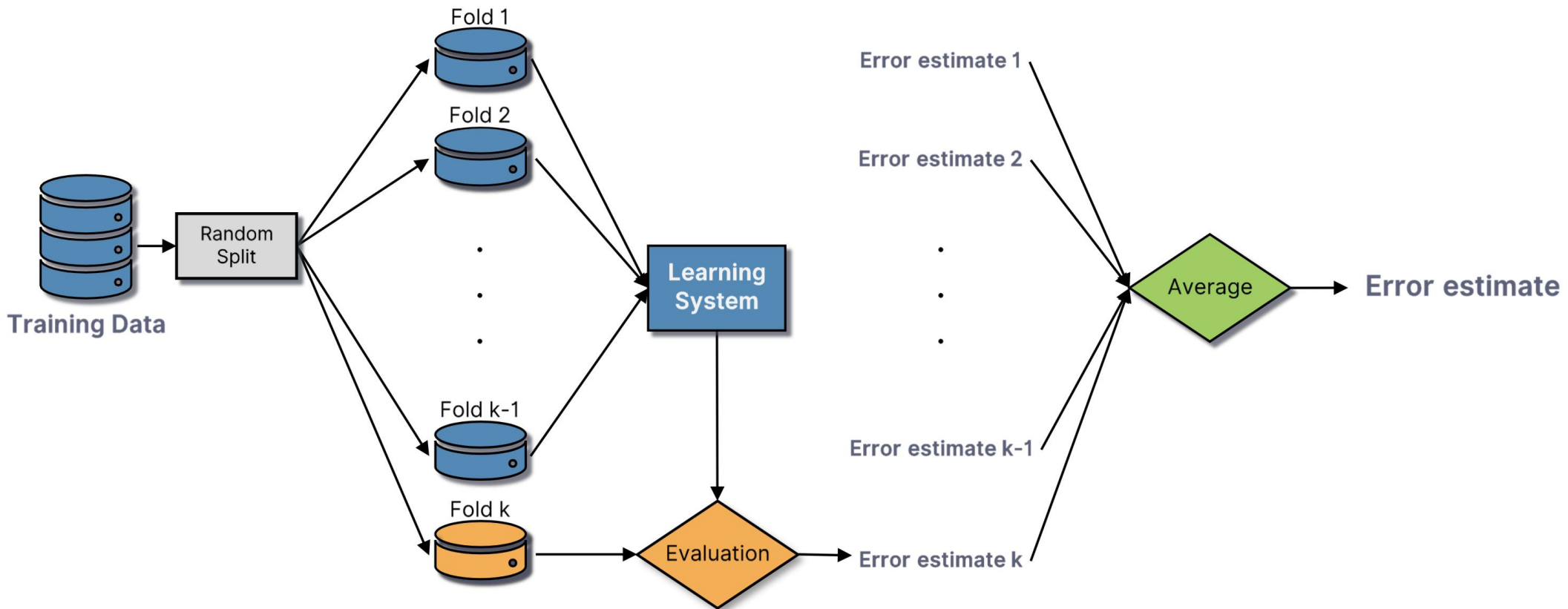
Cross-validation



Cross-validation



Cross-validation



Îmbunătățirea capacității de generalizare

Early stopping

- Oprirea învățării atunci când observăm că eroarea pe validare începe să crească

Regularizare

- Adăugarea unui termen care să penalizeze complexitatea funcției de învățare, impunând restricții de netezire sau limite asupra normei vectorului de ponderi

$$\min_f \sum_{i=1}^n V(f(\hat{x}_i), \hat{y}_i) + \lambda R(f)$$

Evaluare performanței

Cum evaluăm un sistem de învățare automată?

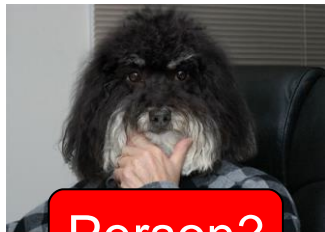
- Măsurăm acuratețea / eroarea pe datele de test:



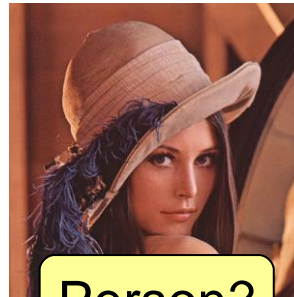
Car?



Person?



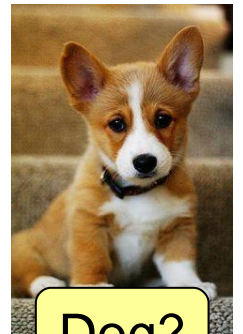
Person?



Person?



Dog?



Dog?

- Acuratețea: 4 corecte din 6 = 66.67%
- Eroarea: 2 greșite din 6 = 33.33%

Cum evaluăm un sistem de învățare automată?

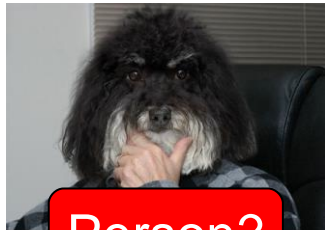
- Construim matricea de confuzie



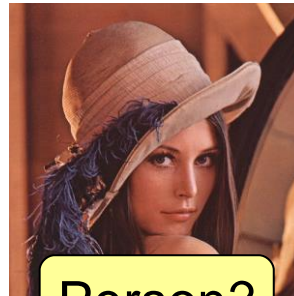
Car?



Person?



Person?



Person?



Dog?



Dog?

- Acuratețea: suma elementelor de pe diagonală principală supra numărul de componente diferite de zero (4/6)
- Eroarea: suma elementelor rămase în afara diagonalei supra numărul de componente diferite de zero (2/6)

Predicted Actual	Car	Dog	Person
Car	1	1	0
Dog	0	1	1
Person	0	0	2

Cum evaluăm un sistem de învățare automată?

- Matricea de confuzie în cazul binar



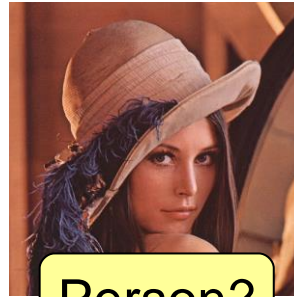
Not?



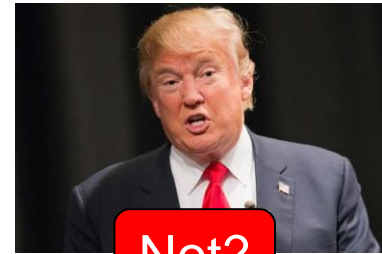
Person?



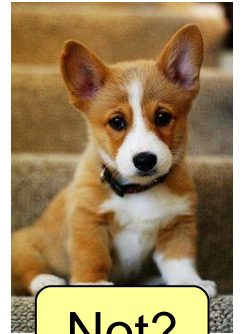
Person?



Person?



Not?



Not?

	Predicted YES	Predicted NO
Actual YES	True Positive	False Negative
Actual NO	False Positive	True Negative

Cum evaluăm un sistem de învățare automată?

- Matricea de confuzie în cazul binar



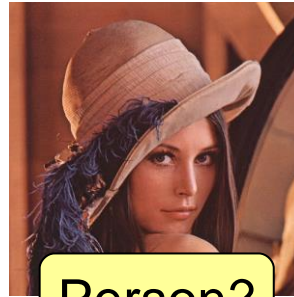
Not?



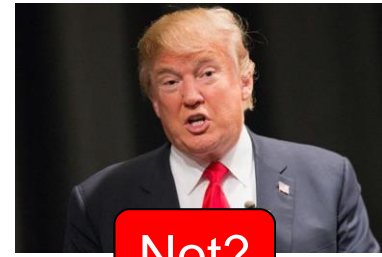
Person?



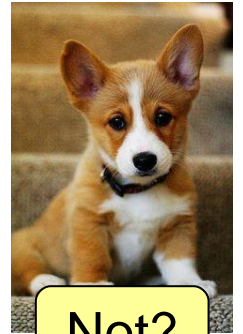
Person?



Person?



Not?



Not?

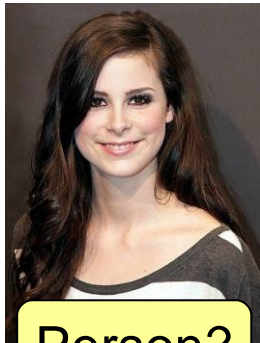
	Predicted YES	Predicted NO
Actual YES	2	False Negative
Actual NO	False Positive	True Negative

Cum evaluăm un sistem de învățare automată?

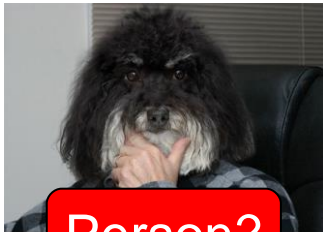
- Matricea de confuzie în cazul binar



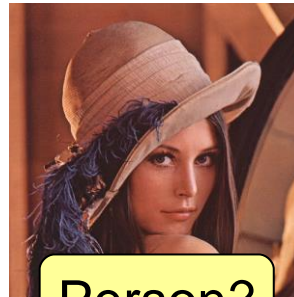
Not?



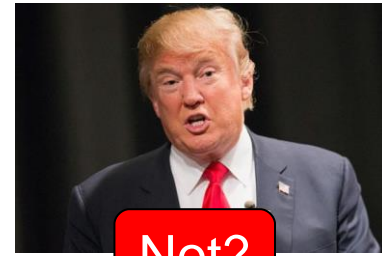
Person?



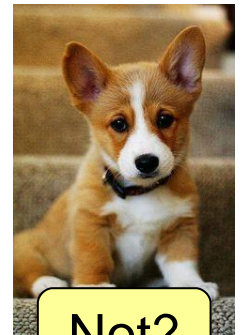
Person?



Person?



Not?



Not?

	Predicted YES	Predicted NO
Actual YES	2	1
Actual NO	False Positive	True Negative

Cum evaluăm un sistem de învățare automată?

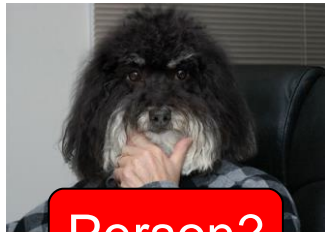
- Matricea de confuzie în cazul binar



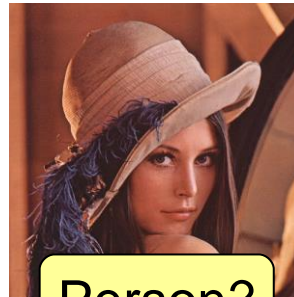
Not?



Person?



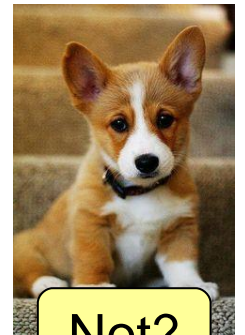
Person?



Person?



Not?



Not?

	Predicted YES	Predicted NO
Actual YES	2	1
Actual NO	1	True Negative

Cum evaluăm un sistem de învățare automată?

- Matricea de confuzie în cazul binar



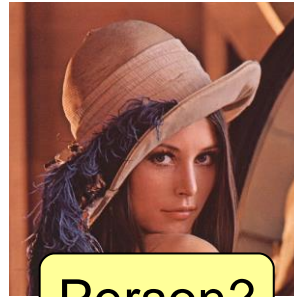
Not?



Person?



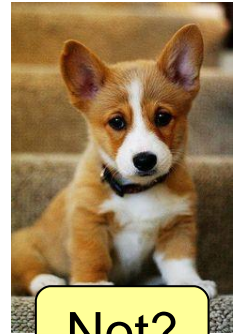
Person?



Person?



Not?



Not?

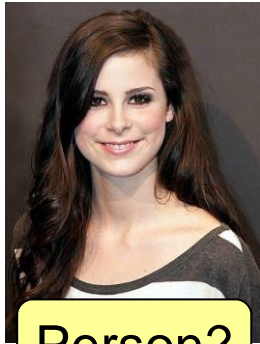
	Predicted YES	Predicted NO
Actual YES	2	1
Actual NO	1	2

Cum evaluăm un sistem de învățare automată?

- Calculul măsurilor Precision și Recall



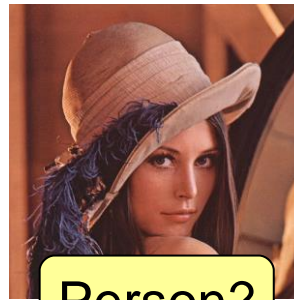
Not?



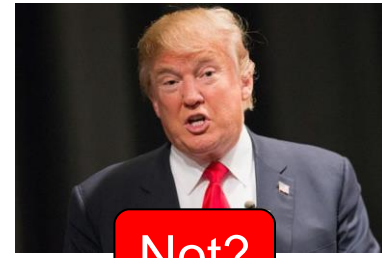
Person?



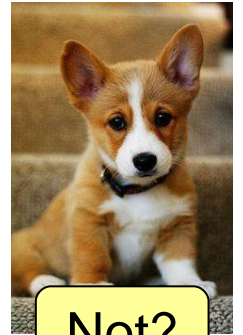
Person?



Person?



Not?



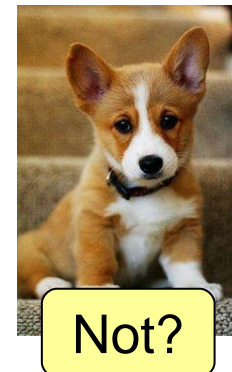
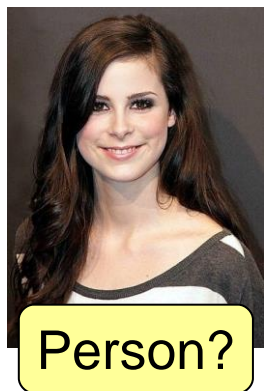
Not?

- $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$
 $= 66.67\%$
- $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$
 $= 66.67\%$

	Predicted YES	Predicted NO
Actual YES	2	1
Actual NO	1	2

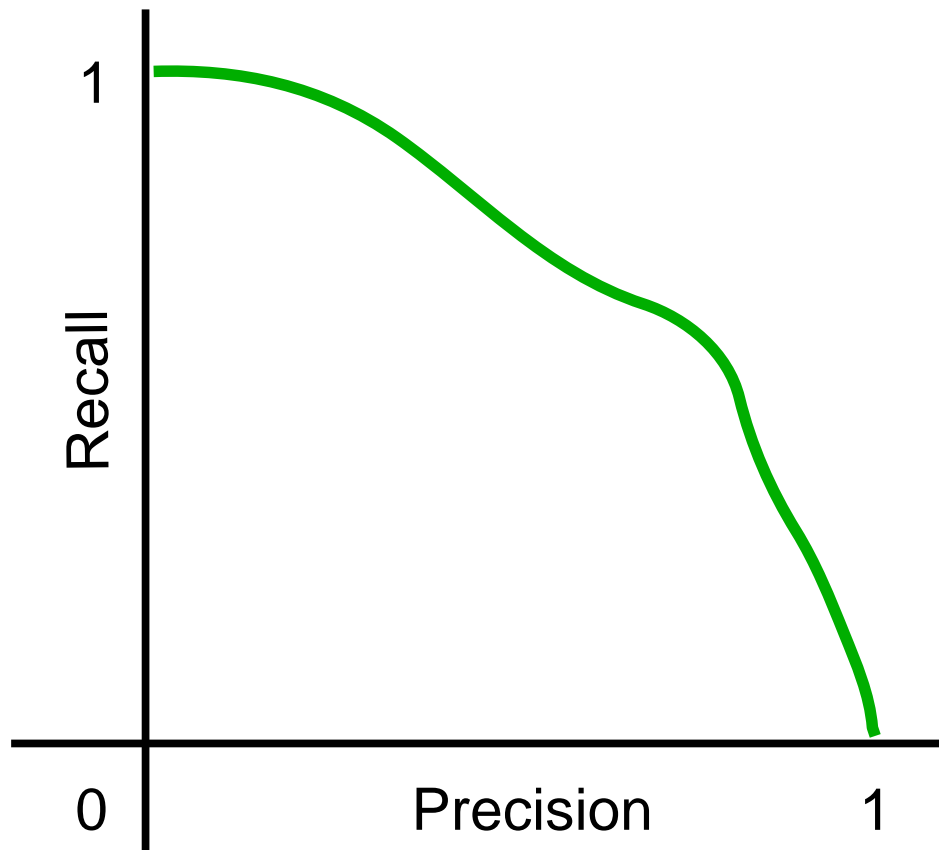
Cum evaluăm un sistem de învățare automată?

- Curba Precision-Recall



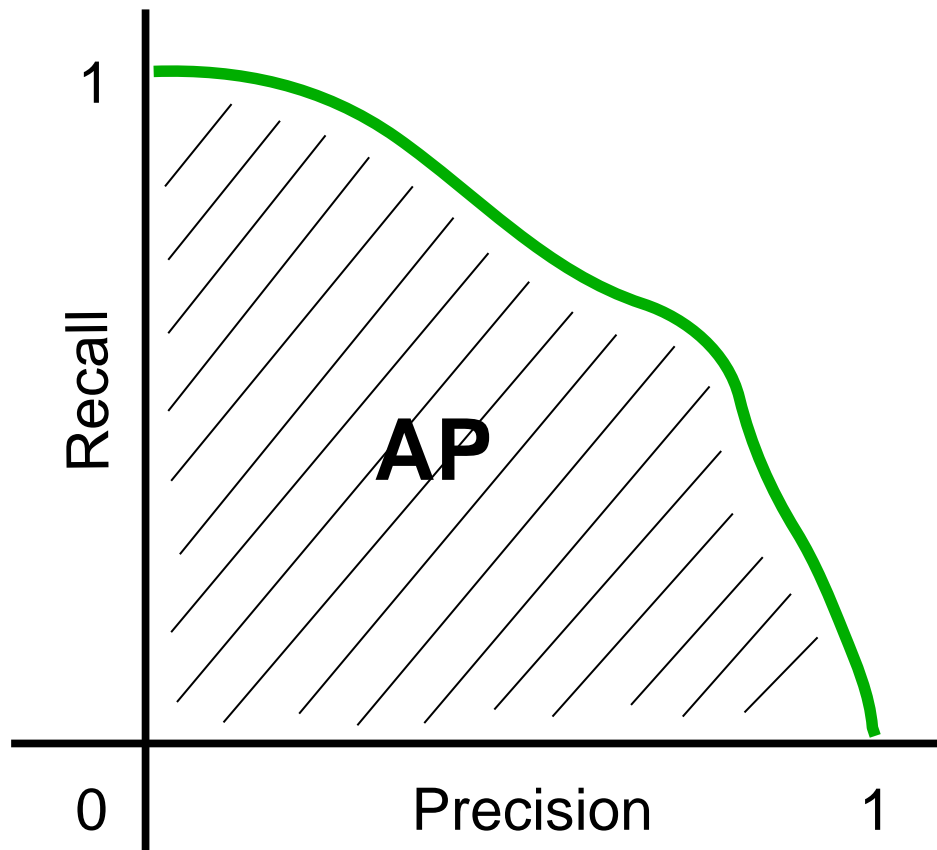
Cum evaluăm un sistem de învățare automată?

- Curba Precision-Recall



Cum evaluăm un sistem de învățare automată?

- Average Precision



Cum evaluăm un sistem de învățare automată?

- Calculul măsurilor TPR și FPR



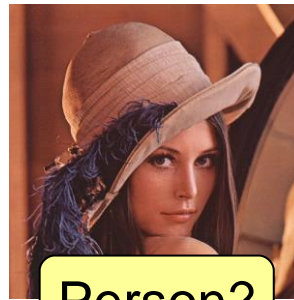
Not?



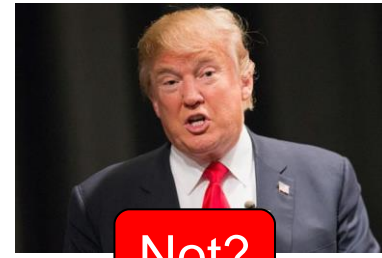
Person?



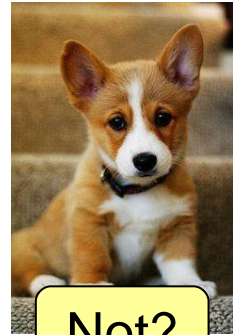
Person?



Person?



Not?



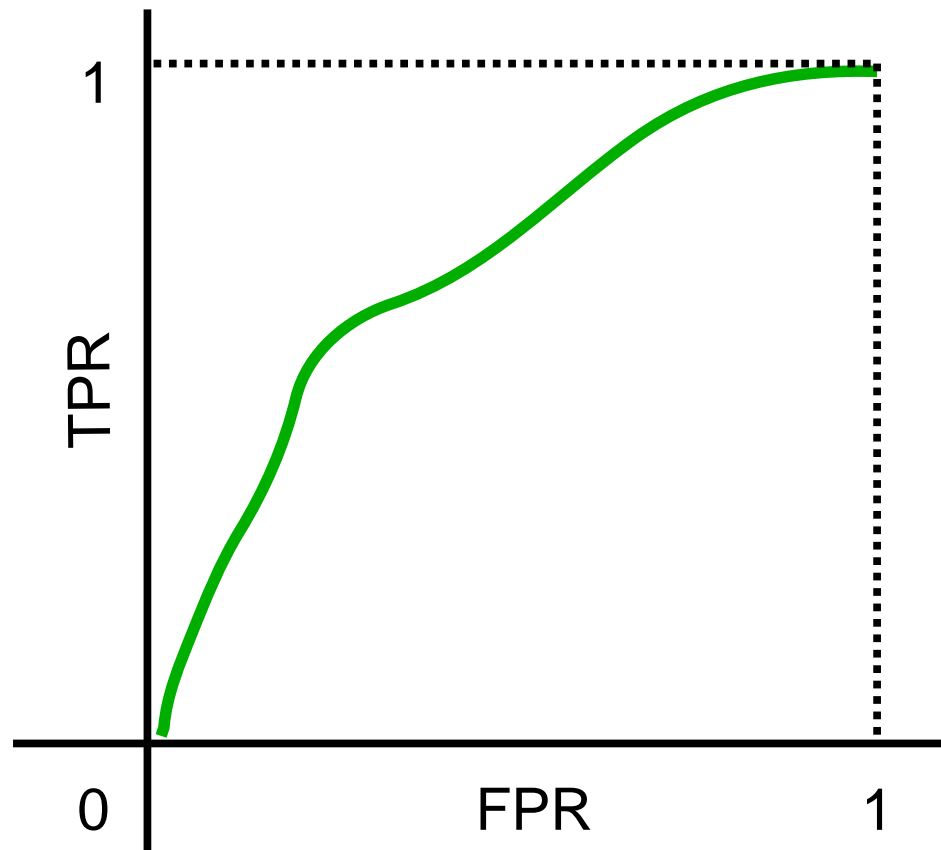
Not?

- $TPR = TP / (TP + FP)$
= 66.67%
- $FPR = FP / (FP + TN)$
= 33.33%

	Predicted YES	Predicted NO
Actual YES	2	1
Actual NO	1	2

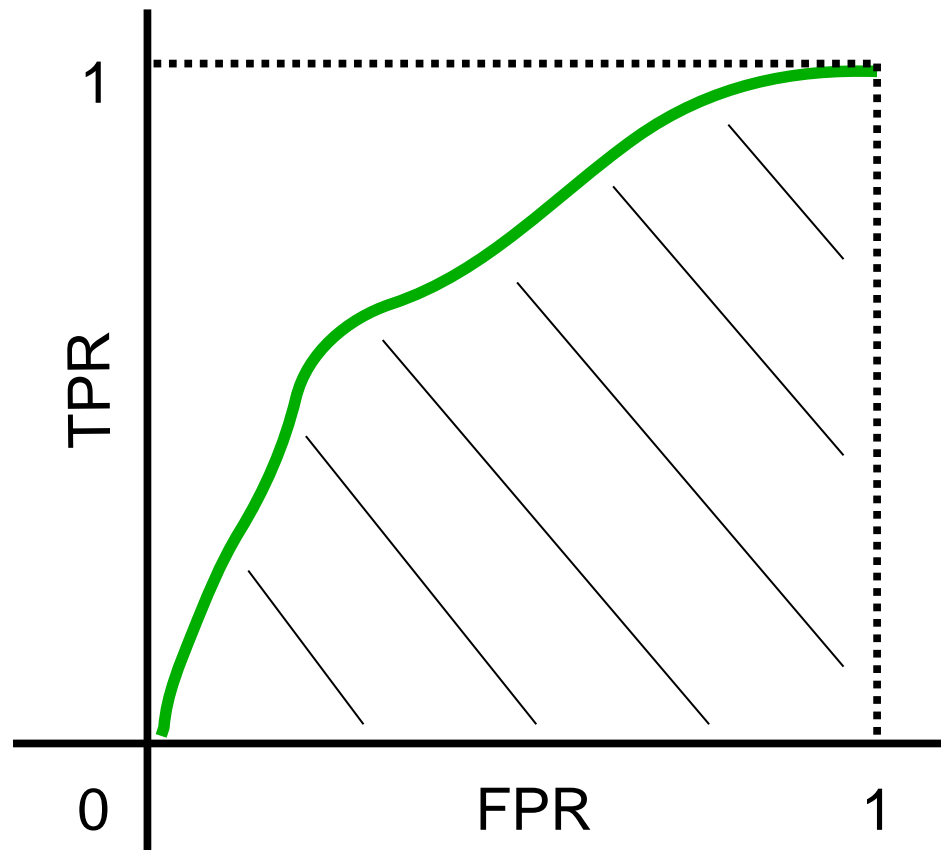
Cum evaluăm un sistem de învățare automată?

- Curba ROC (Receiver Operating Characteristic)



Cum evaluăm un sistem de învățare automată?

- Măsura AUC: Aria de sub curba ROC



Cum evaluăm un sistem de învățare automată?

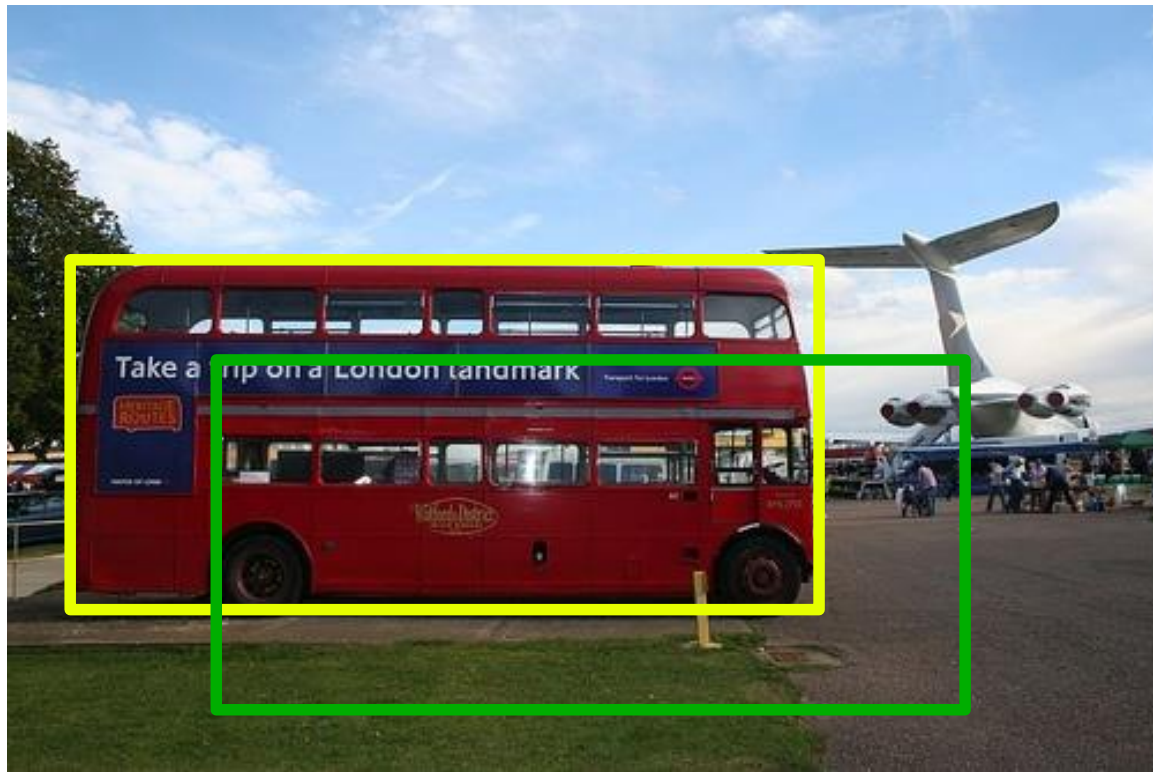
- Măsura F_β

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

- Măsura F_1 este poate cea mai folosită măsură de tipul F_β

Cum evaluăm un sistem de detecție de obiecte?

- Intersecție supra Uniune (indexul Jaccard)



Cum evaluăm un sistem de detecție de obiecte?

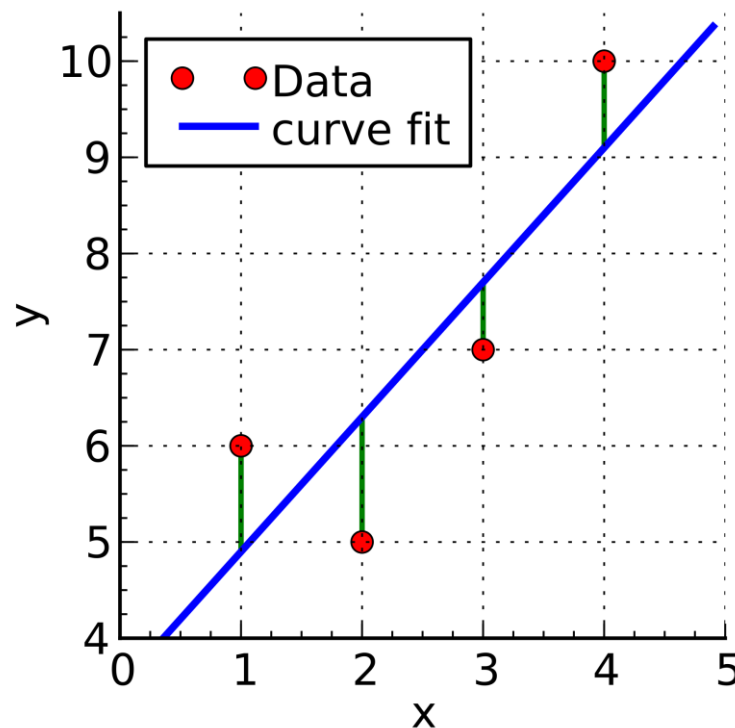
- Intersecție supra Uniune (indexul Jaccard)
- Detecție corectă dacă $J(A,B) > 0.5$



Cum evaluăm un sistem de regresie?

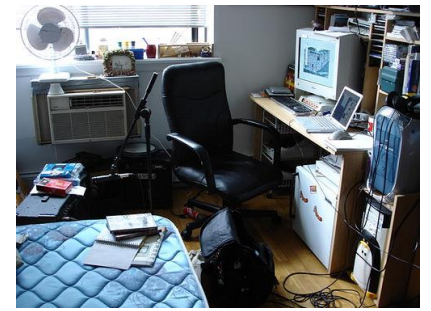
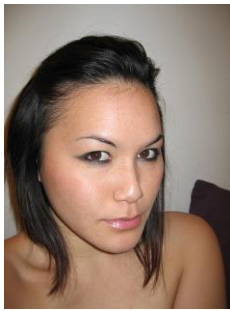
- Media pătratelor erorilor (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$



Cum evaluăm un sistem de regresie?

- Ordinea dificultății conform oamenilor



- Ordinea dificultății prezisă de sistem



Cum evaluăm un sistem de regresie?

- Corelația Kendall Tau:

$$\tau_a = \frac{P - Q}{\frac{n(n-1)}{2}}$$

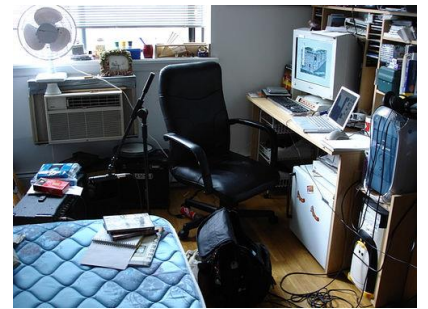
- Măsură ordinală bazată pe perechi concordante (P) și discordante (Q)

$$P = |\{(i, j) : 1 \leq i < j \leq n, (x_i - x_j)(y_i - y_j) > 0\}|$$

$$Q = |\{(i, j) : 1 \leq i < j \leq n, (x_i - x_j)(y_i - y_j) < 0\}|$$

Cum evaluăm un sistem de regresie?

- Ordinea dificultății conform oamenilor

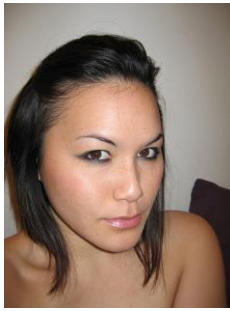


- Concordantă cu ordinea prezisă de sistem



Cum evaluăm un sistem de regresie?

- Ordinea dificultății conform oamenilor

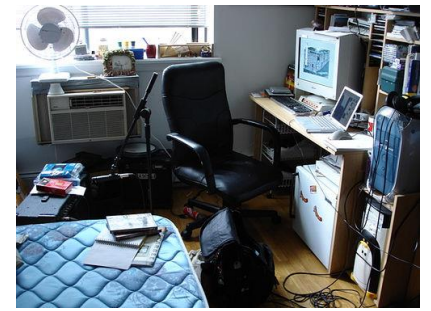
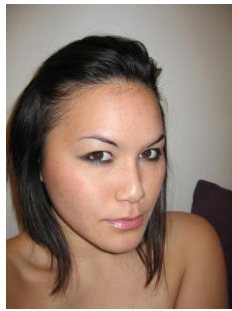


- Discordantă cu ordinea prezisă de sistem



Cum evaluăm un sistem de regresie?

- Cât este corelația Kendall Tau?

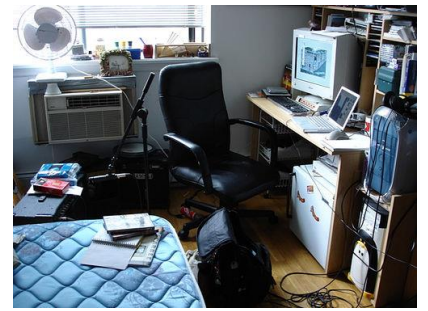
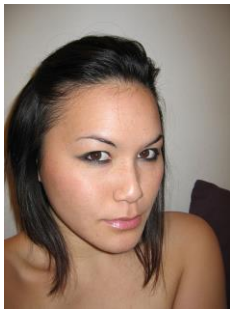


- $P = ?$, $Q = ?$



Cum evaluăm un sistem de regresie?

- Cât este corelația Kendall Tau?



- $P = 8$, $Q = 2$, Kendall Tau = $(8-2) / 10 = 0.6$

