

# KLASIFIKACIJA

**JELENA JOVANOVIĆ**

Email: [jeljov@gmail.com](mailto:jeljov@gmail.com)

Web: <http://jelenajovanovic.net>

# PREGLED PREDAVANJA

- Šta je klasifikacija?
- Binarna i više-klasna klasifikacija
- Algoritmi klasifikacije
- Mere uspešnosti klasifikatora

# ŠTA JE KLASIFIKACIJA?

- Zadatak određivanja klase kojoj neka instanca pripada
  - instanca je opisana vrednošću atributa;
  - skup mogućih klasa je poznat i dat
- Klase su date kao nominalne vrednosti, npr.
  - klasifikacija email poruka: spam, not-spam
  - klasifikacija novinskih članaka: politika, sport, kultura i sl.

# BINARNA I VIŠE-KLASNA KLASIFIKACIJA

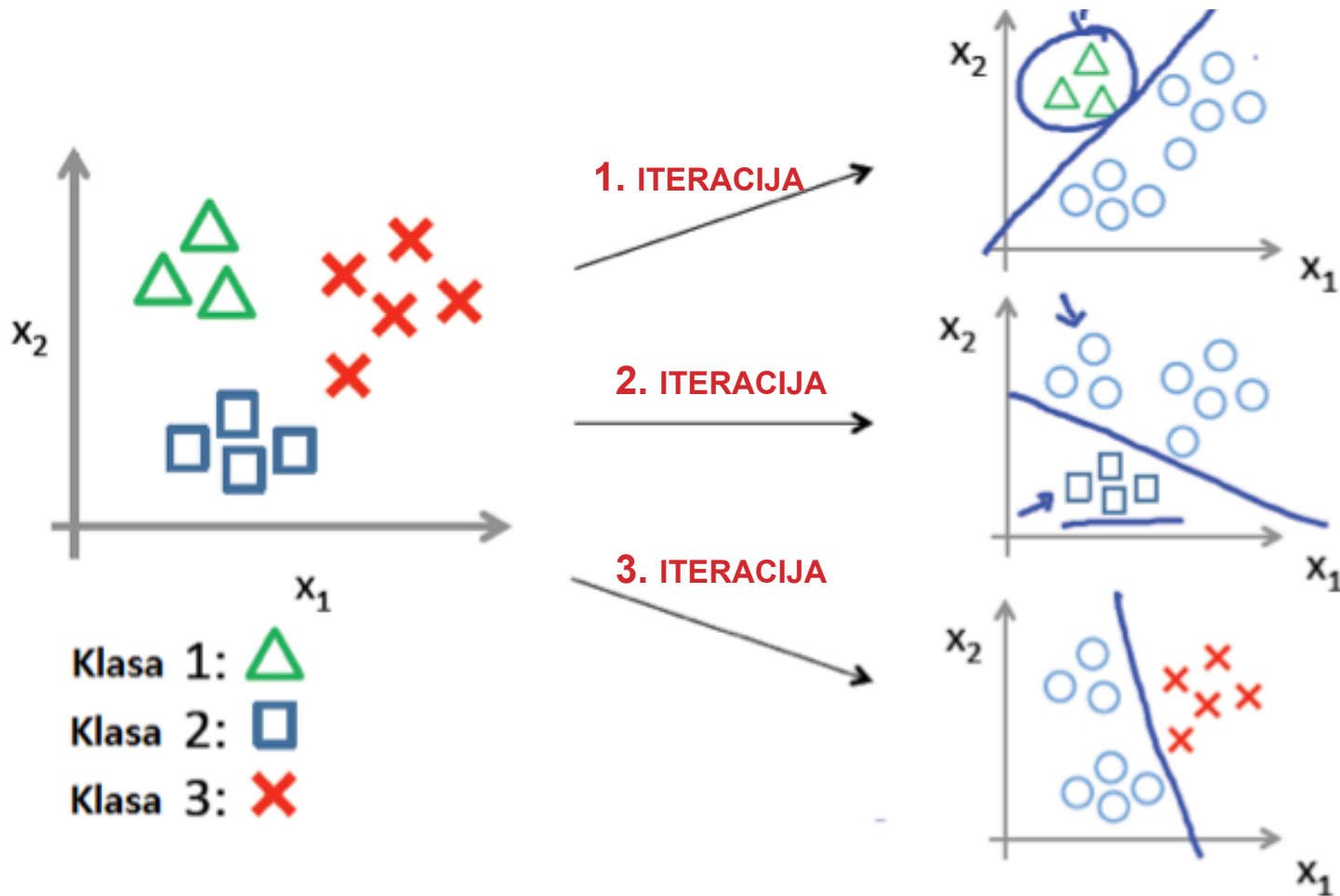
Zavisno od broja klasa, razlikujemo:

- *binarnu* klasifikaciju - postoje dve klase
- *više-klasnu* klasifikacija - postoji više klasa u koje instance treba razvrstati

Princip rada algoritma u oba slučaja je gotovo isti:

u slučaju postojanja više klasa, algoritam iterativno uči, tako da u svakoj iteraciji “nauči” da jednu od klasa razgraniči od svih ostalih

# VIŠE-KLASNA KLASIFIKACIJA



# ALGORITMI KLASIFIKACIJE

Postoje brojni pristupi/algoritmi za klasifikaciju:

- Logistička regresija
- Naïve Bayes
- Algoritmi iz grupe Stabala odlučivanja
- Algoritmi iz grupe Neuronskih mreža
- k-Nearest Neighbor (kNN)
- Support Vector Machines (SVM)
- ...

# MERE USPEŠNOSTI KLASIFIKATORA

Neke od najčešće korišćenih metrika:

- Matrica zabune (Confusion Matrix)
- Tačnost (Accuracy)
- Preciznost (Precision) i Odziv (Recall)
- F mera (F measure)
- Površina ispod ROC krive (Area Under the Curve - AUC)

# MATRICA ZABUNE (CONFUSION MATRIX)

Služi kao osnova za računanje mera performansi (uspešnosti) algoritama klasifikacije

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

TP = True Positive

FP = False Positive

TN = True Negative

FN = False Negative



# TAČNOST (ACCURACY)

Tačnost (Accuracy) predstavlja procenat slučajeva (instanci) koji su uspešno (korektno) klasifikovani

$$\text{Accuracy} = (TP + TN) / N$$

gde je:

- TP – True Positive; TN – True Negative
- N – ukupan broj uzoraka (instanci) u skupu podataka

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

# TAČNOST (ACCURACY)

U slučaju vrlo neravnomerne raspodele instanci između klasa (tzv. skewed classes), ova mera je nepouzdana

Npr. u slučaju klasifikacije poruka na spam vs. not-spam, možemo imati skup za trening sa 0.5% spam poruka

Ako primenimo “klasifikator” koji svaku poruku svrstava u not-spam klasu, dobijamo tačnost od 99.5%

Očigledno je da ova metrika nije pouzdana i da su u slučaju skewed classes potrebne druge metrike

# PRECIZNOST (PRECISION) I ODZIV (RECALL)

**Precision** = TP / no. predicted positive = TP / (TP + FP)

Npr. od svih poruka koje su *označene kao spam* poruke, koji procenat čine poruke koje su stvarno spam

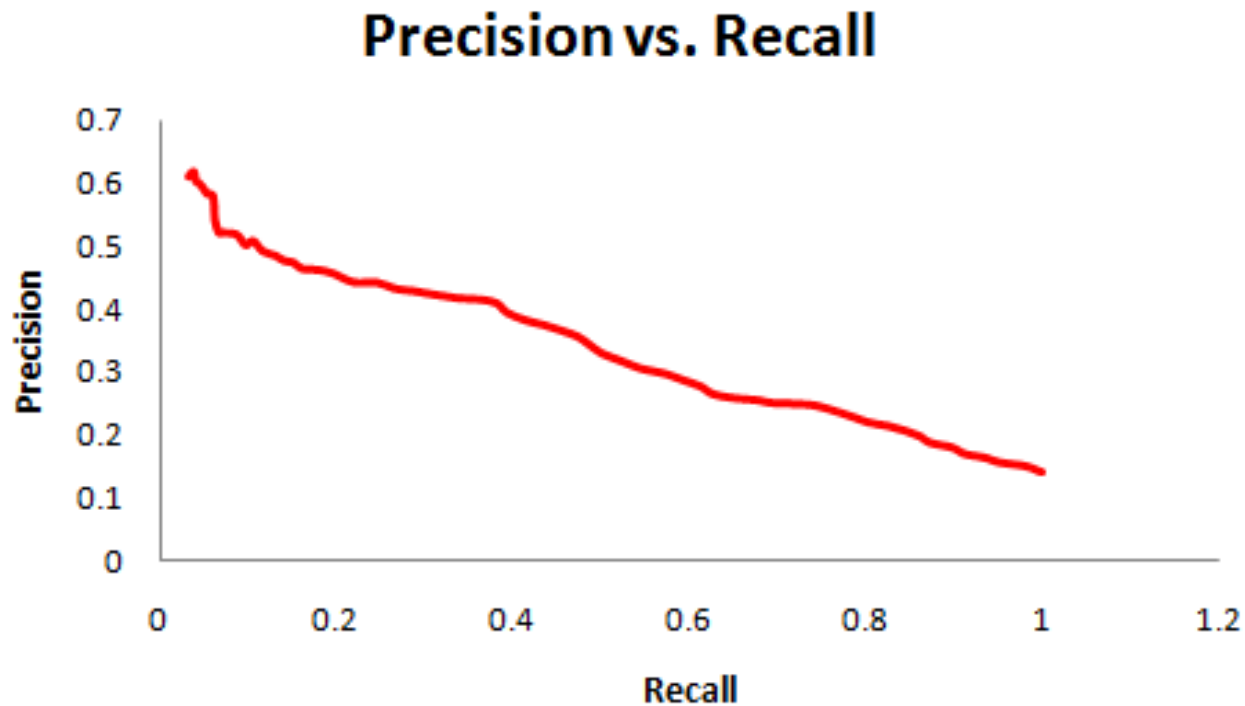
**Recall** = TP / no. actual positive = TP / (TP + FN)

Npr. od svih poruka koje su *stvarno spam* poruke, koji procenat poruka je detektovan/klasifikovan kao spam

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

# PRECIZNOST I ODZIV

U praksi je nužno praviti kompromis između ove dve mere: ako želimo da povećamo Odziv, smanjićemo Preciznost, i obrnuto.



# F MERA (F MEASURE)

F mera kombinuje Preciznost i Odziv i omogućuje jednostavnije poređenje dva ili više algoritama

$$F = (1 + \beta^2) * \text{Precision} * \text{Recall} / (\beta^2 * \text{Precision} + \text{Recall})$$

Parametar  $\beta$  kontroliše koliko više značaja će se pridavati Odzivu u odnosu na Preciznost

U praksi se najčešće koristi tzv. F1 mera („balansirana“ F mera) koja daje podjednak značaj i Preciznosti i Odzivu:

$$F1 = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

# POVRŠINA ISPOD ROC KRIVE

Površina ispod ROC\* krive – Area Under the Curve (AUC):

- meri diskriminacionu moć klasifikatora tj. sposobnost da razlikuje instance koje pripadaju različitim klasama
- primenjuje se za merenje performansi binarnih klasifikatora
- vrednost za AUC se kreće u intervalu 0-1
- za metodu slučajnog izbora važi da je  $AUC = 0.5$ ; što je AUC vrednost klasifikatora  $> 0.5$ , to je klasifikator bolji
  - 0.7–0.8 se smatra prihvatljivim; 0.8–0.9 jako dobrim; sve  $> 0.9$  je odlično

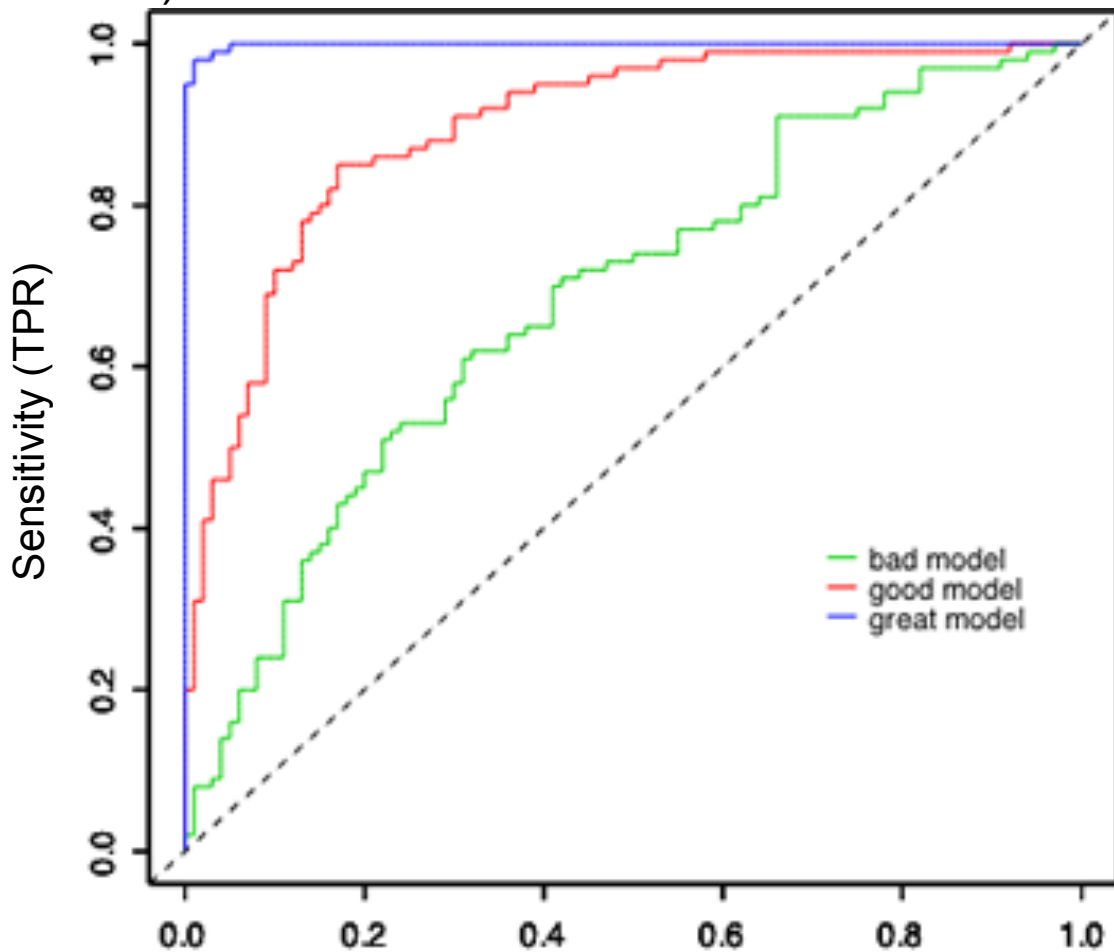
\*ROC = Receiver Operating Characteristic;

[http://en.wikipedia.org/wiki/Receiver\\_operating\\_characteristic](http://en.wikipedia.org/wiki/Receiver_operating_characteristic)

# POVRŠINA ISPOD ROC KRIVE

**Sensitivity ili True Positive Rate (TPR)**

$$\text{TPR} = \text{TP} / (\text{TP} + \text{FN})$$



1 - Specificity (TNR)

**True Negative Rate (TNR)**

$$\text{TNR} = \text{TN} / (\text{FP} + \text{TN})$$

# ZAHVALNICE I PREPORUKE



# PREPORUKE I ZAHVALNICE

## MACHINE LEARNING @ STANFORD

- Coursera: <https://www.coursera.org/course/ml>
- Stanford YouTube channel:  
[http://www.youtube.com/view\\_play\\_list?p=A89DCFA6ADACE599](http://www.youtube.com/view_play_list?p=A89DCFA6ADACE599)

# PREPORUKA

- [article] Visual Introduction to Machine Learning:  
<http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>
- [blog post] Choosing a Machine Learning Classifier:  
<http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/>
- [web site] Machine Learning Classifiers Gallery:  
[http://home.comcast.net/~tom.fawcett/public\\_html/ML-gallery/pages/](http://home.comcast.net/~tom.fawcett/public_html/ML-gallery/pages/)
- [article] IU scientists use Instagram data to forecast top models at New York Fashion Week (<http://goo.gl/ovepjax>)
- [podcast] Data Stories podcast #27; topic: “Big Data Skepticism” (<http://goo.gl/KKPGuW>)
  - the podcast mentioned a study that was aimed at the prediction of demographic characteristics of Facebook users based on their Likes (<http://goo.gl/fykOyt>)

(Anonimni) upitnik za vaše kritike,  
komentare, predloge:

<http://goo.gl/cqdp3l>