Multiclass Classification
Principal Component Analysis

Lecture 06

Recap

Lecture 5:
Linear Classification

- Linear classification
    - margin
    - loss functions
- Logistic regression
    - sigmoid derivation
    - Maximum Likelihood Estimation
    - logistic loss
- Metrics in classification
    - Accuracy, Balanced accuracy
    - Precision, Recall, F-score
    - ROC curve, PR curve, AUC

Outline

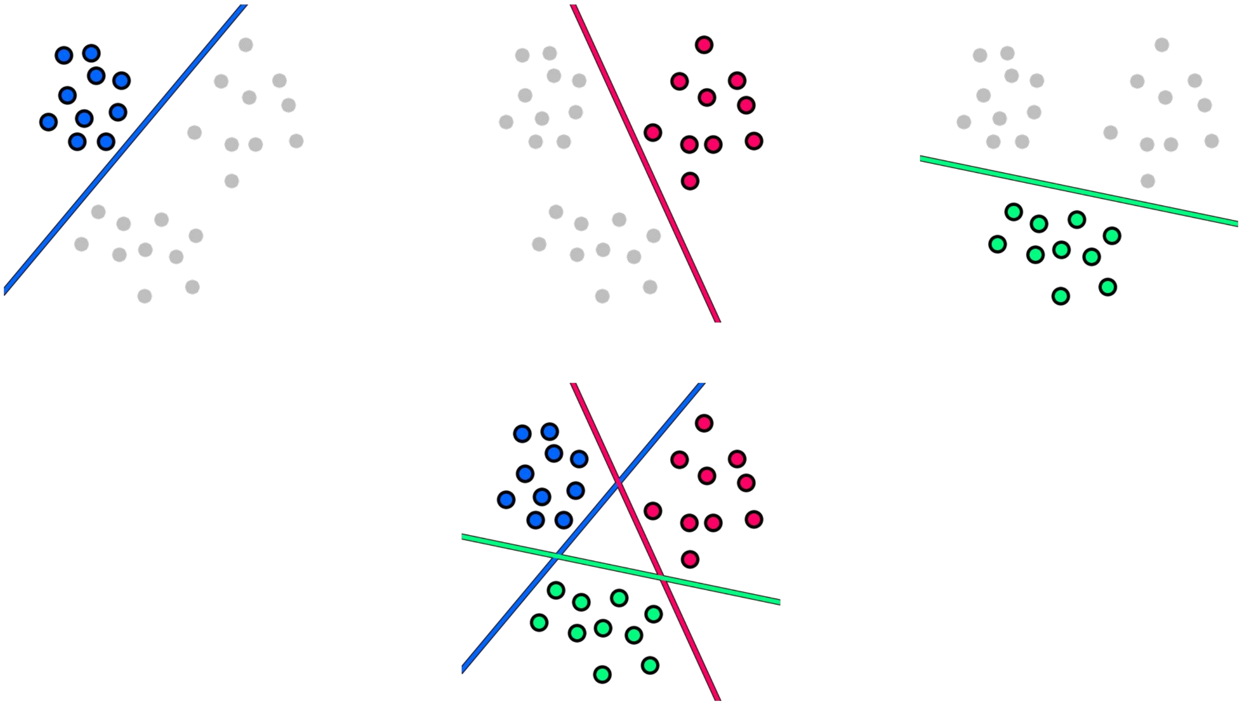
- Multiclass aggregation strategies
  - One vs Rest
  - One vs One
- Metrics in classification (again):
  - Precision, Recall, F-score
  - ROC curve, PR curve, AUC
  - Confusion matrix
- Dimensionality reduction and PCA
  - Connections with SVD

Multiclass aggregation strategies

girafe
ai

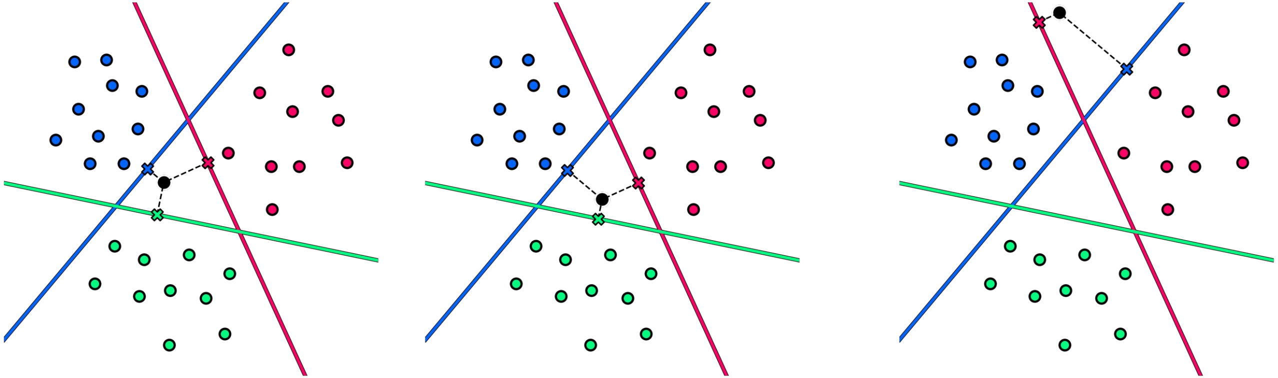
01

One vs Rest

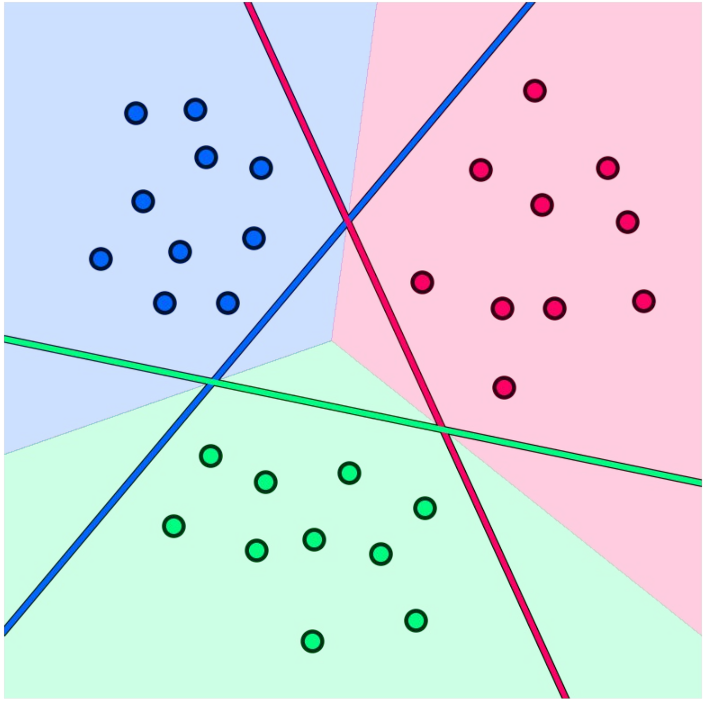


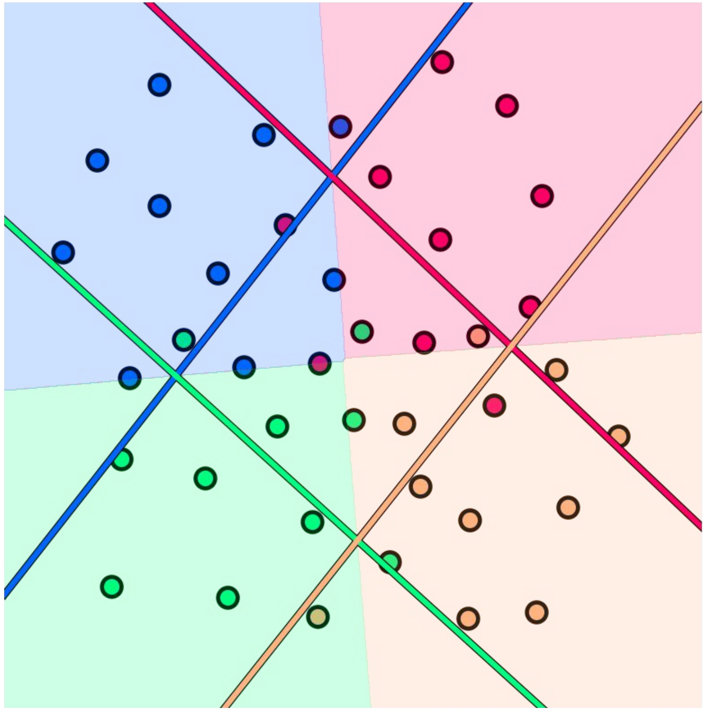
Images source

One vs Rest: unclassified regions

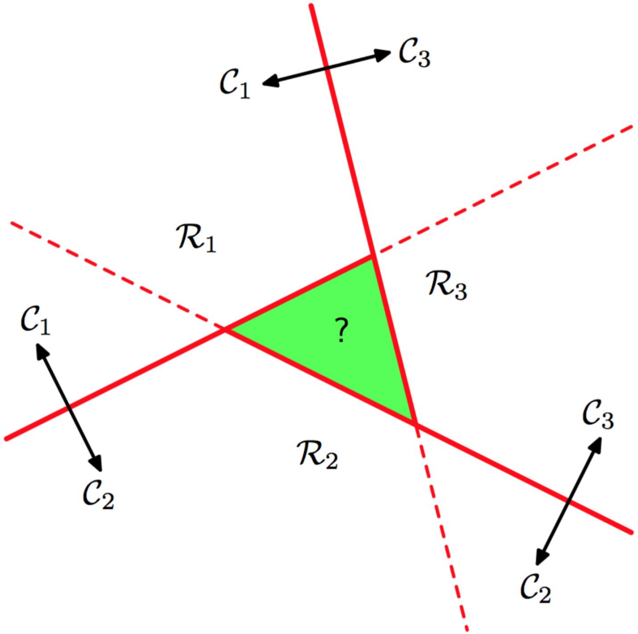


One vs Rest: final result



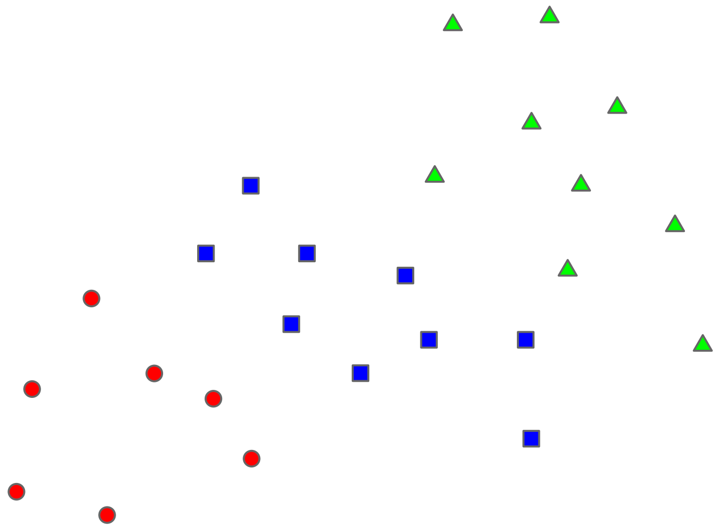


One vs One



$\mathcal{C}_1$
$\mathcal{C}_3$
$\mathcal{R}_1$
$\mathcal{R}_3$
$\mathcal{C}_1$
?
$\mathcal{R}_2$
$\mathcal{C}_3$
$\mathcal{C}_2$
$\mathcal{C}_2$

Failure case?



Summary

	One vs Rest	One vs One
#classifiers	k	k(k-1)/2
dataset for each	full	subsampled

Metrics: Multi-class Scenario

girafe
ai

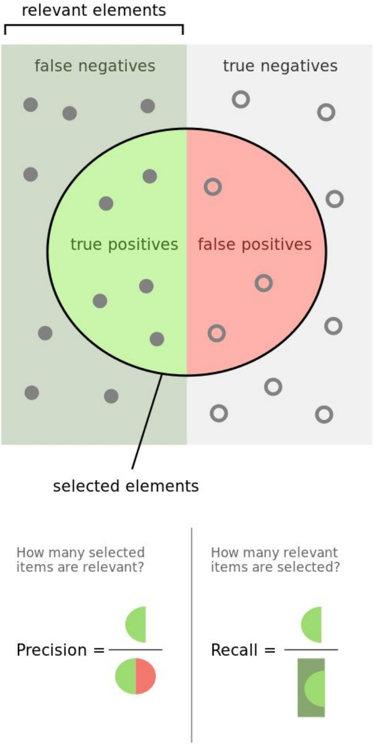
02

Metrics

- Multiclass generalization:
    - Precision
    - Recall
    - F-score
    - ROC-AUC
    - PR-AUC
- Confusion matrix

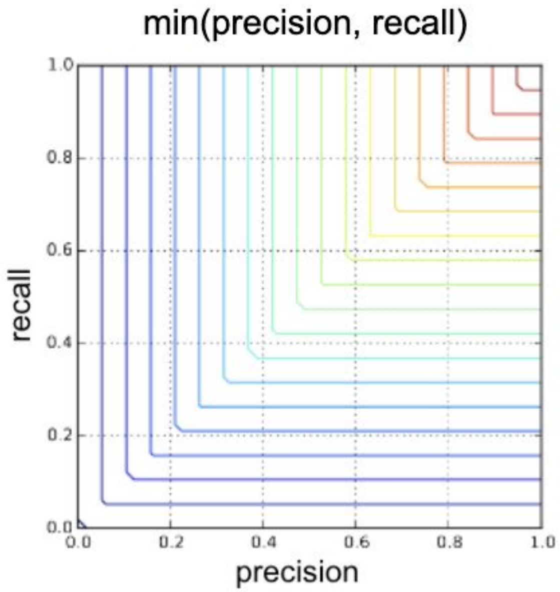
Precision and Recall

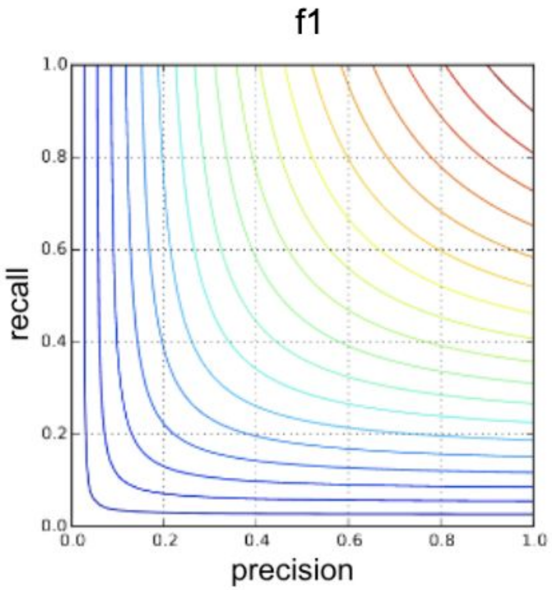
		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**



$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

F-score motivation





F-score

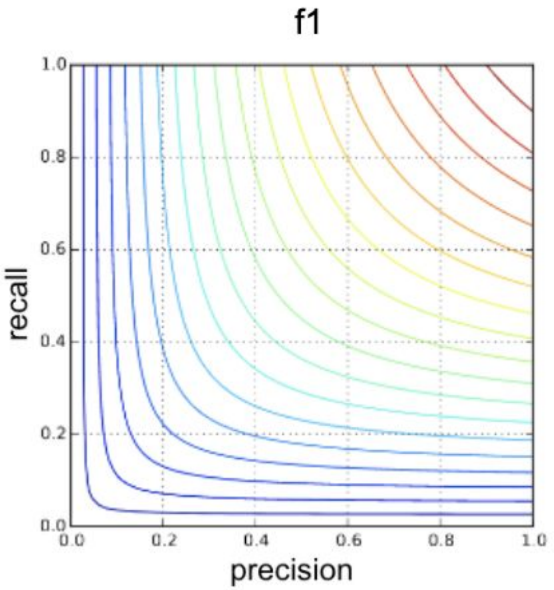
Harmonic mean of precision and recall

Closer to smaller one

$$F_1 = \frac{2}{\mathrm{precision}^{-1} + \mathrm{recall}^{-1}} = 2\,\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

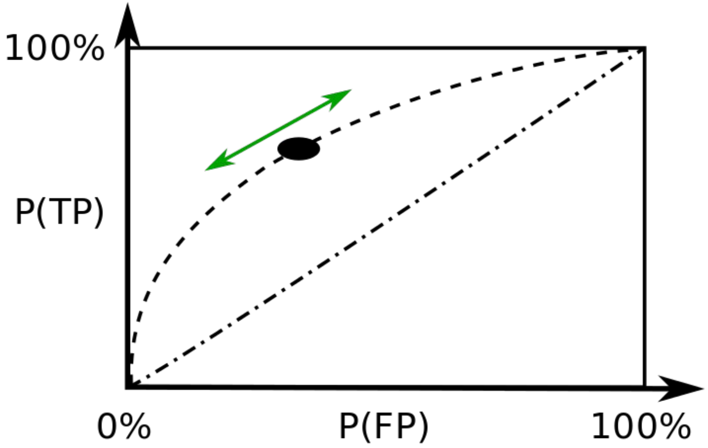
Generalization to different ratio between Precision and Recall

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



Receiver Operating Characteristic (ROC)

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**



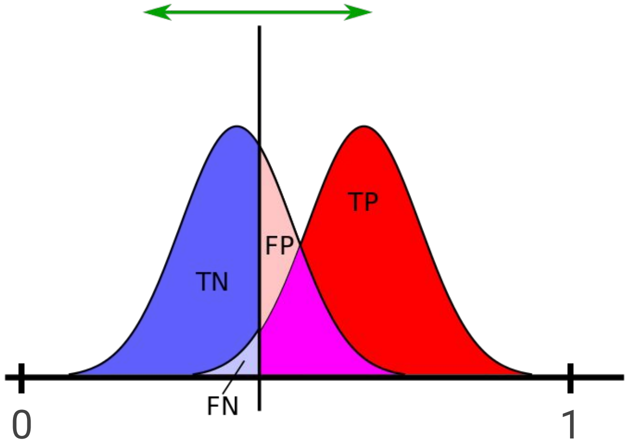
$$\mathrm{FPR} = \frac{FP}{FP+TN}$$

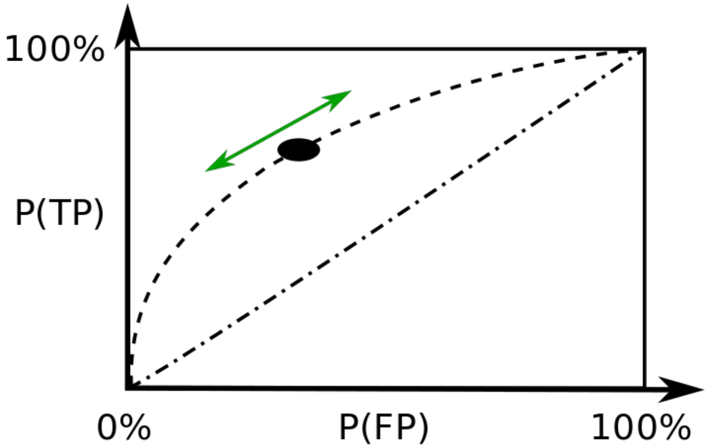
$$\mathrm{TPR} = \frac{TP}{TP+FN} (= \mathrm{Recall})$$

Receiver Operating Characteristic (ROC)

Classifier needs to predict probabilities

Objects get sorted by positive probability





Line is plotted as threshold moves

Receiver Operating Characteristic (ROC)

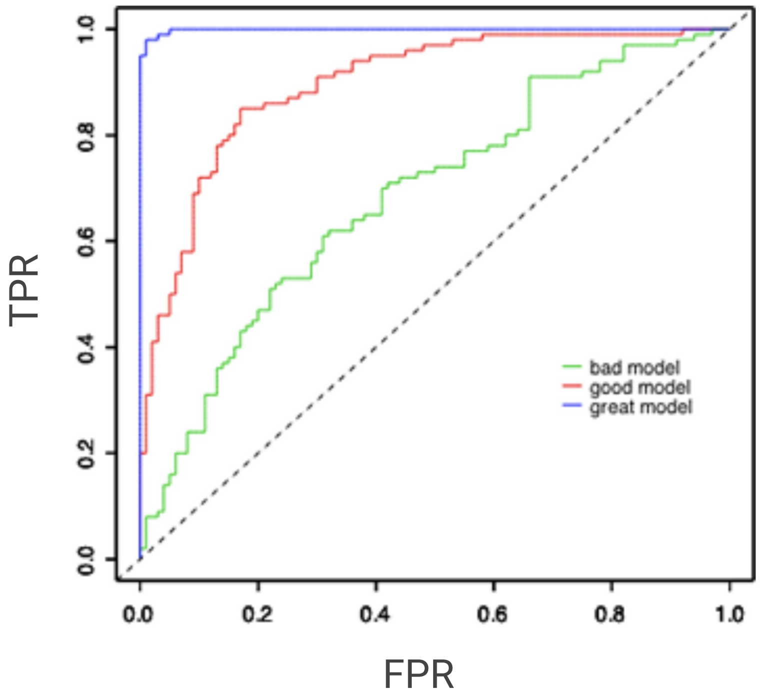
Baseline is random predictions

Always above diagonal (for reasonable classifier)

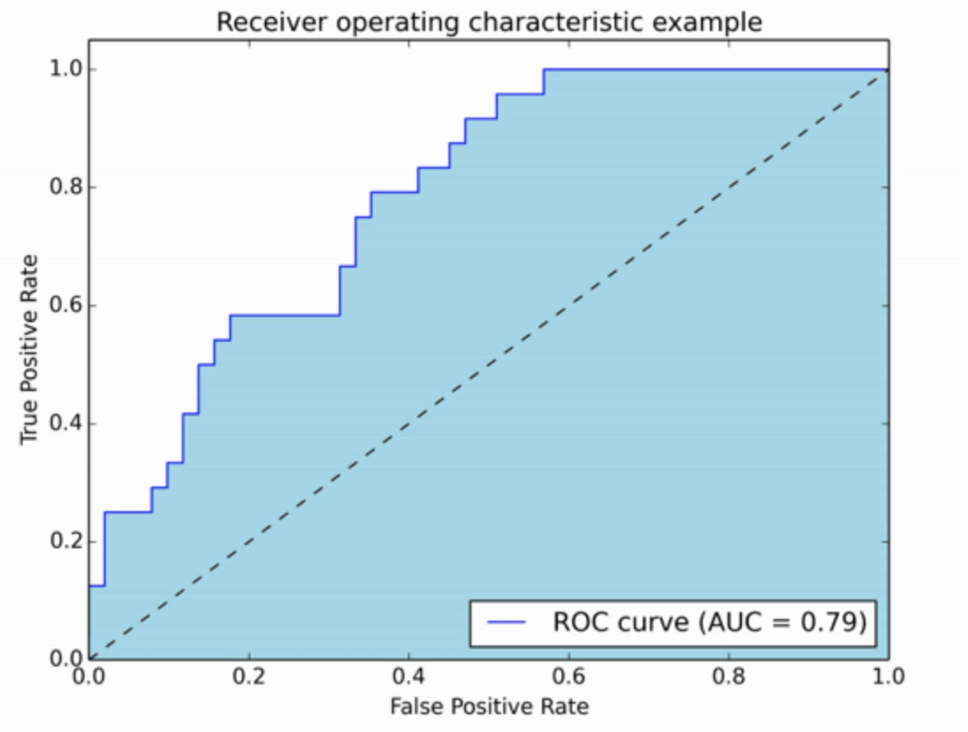
If below - change sign of predictions

Strictly higher curve means better classifier

Number of steps (thresholds) not bigger than dataset



ROC Area Under Curve (ROC-AUC)



Effectively lays in (0.5, 1)

Bigger ROC-AUC doesn't imply

higher curve everywhere

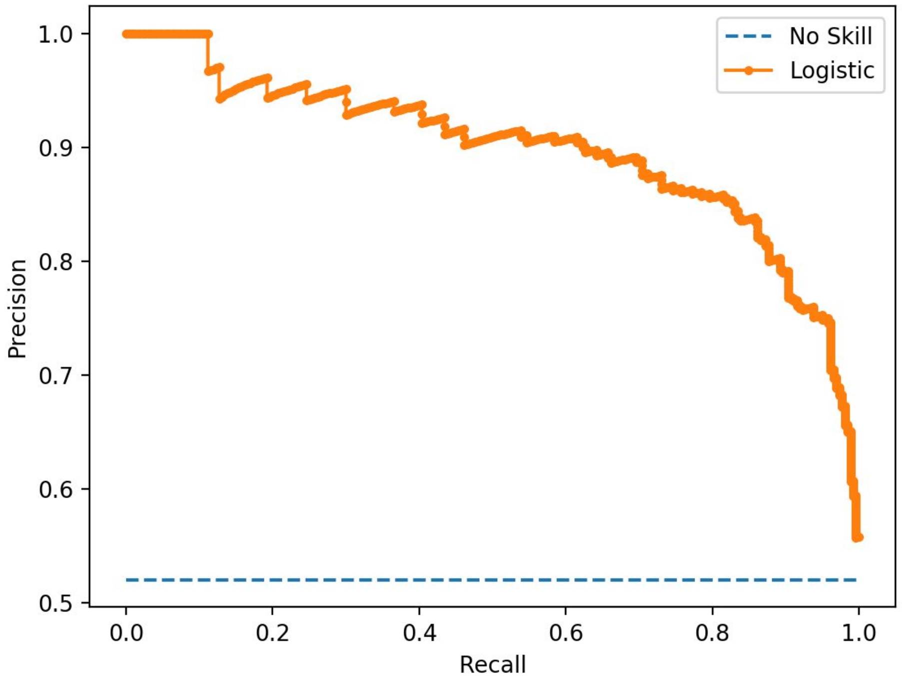
[More explanations with pictures]

Precision-Recall Curve

AUC is in (0, 1)

Source of AP metric
(important for next semester)

Nice article



Multiclass metrics

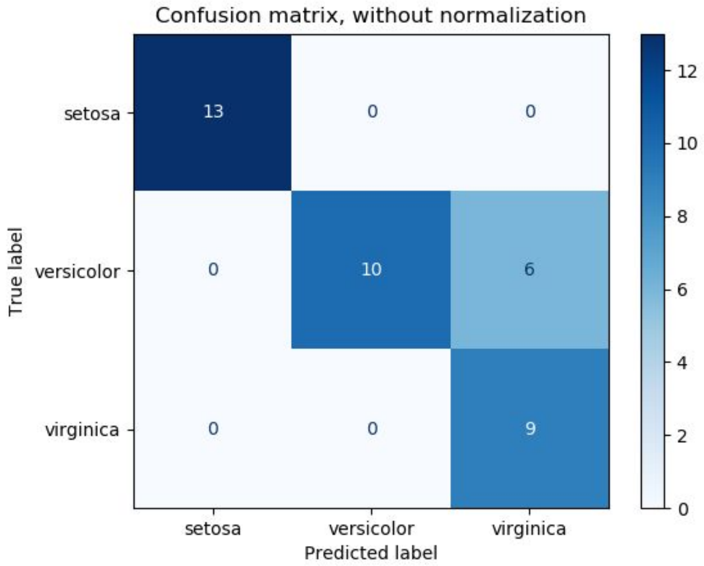
As with linear models we need some magic to measure multiclass problems

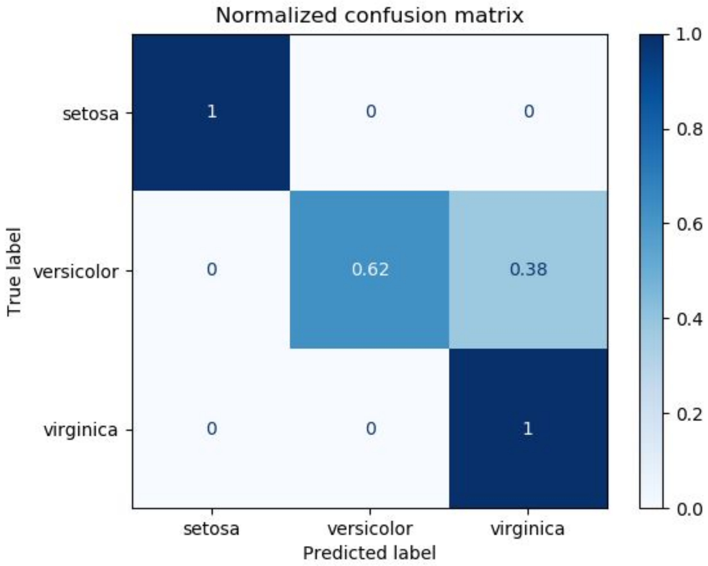
Basically it's mean of one or another kind

Detailed info here and here

average	Precision	Recall	F_beta
"micro"	$P(y, \hat{y})$	$R(y, \hat{y})$	$F_\beta(y, \hat{y})$
"samples"	$\frac{1}{\|S\|} \sum_{s \in S} P(y_s, \hat{y}_s)$	$\frac{1}{\|S\|} \sum_{s \in S} R(y_s, \hat{y}_s)$	$\frac{1}{\|S\|} \sum_{s \in S} F_\beta(y_s, \hat{y}_s)$
"macro"	$\frac{1}{\|L\|} \sum_{l \in L} P(y_l, \hat{y}_l)$	$\frac{1}{\|L\|} \sum_{l \in L} R(y_l, \hat{y}_l)$	$\frac{1}{\|L\|} \sum_{l \in L} F_\beta(y_l, \hat{y}_l)$
"weighted"	$\frac{1}{\sum_{l \in L} \|\hat{y}_l\|} \sum_{l \in L} \|\hat{y}_l\| P(y_l, \hat{y}_l)$	$\frac{1}{\sum_{l \in L} \|\hat{y}_l\|} \sum_{l \in L} \|\hat{y}_l\| R(y_l, \hat{y}_l)$	$\frac{1}{\sum_{l \in L} \|\hat{y}_l\|} \sum_{l \in L} \|\hat{y}_l\| F_\beta(y_l, \hat{y}_l)$

Confusion matrix





Principal Component Analysis

girafe
ai

02

Dimensionality Reduction

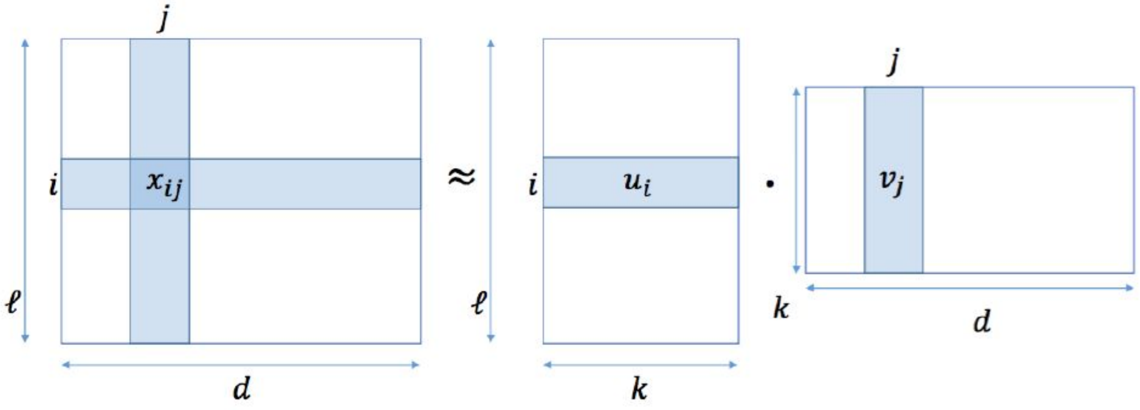
- In ML we often work with high-dimensional data
  - Hundreds or thousands of features

- Hard to visualize
- Slow training
- Some models perform worse on high-dimensional sparse input

Dimensionality Reduction

- Factorization into smaller-rank matrices

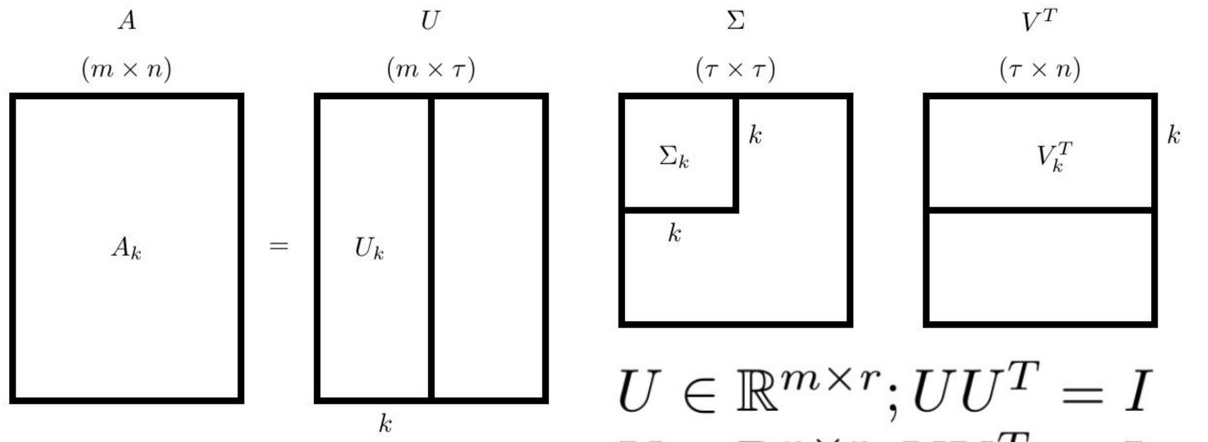
$$X_{l,d} \approx U_{l,k} \cdot V^T_{k,d} \qquad ||X - UV^T|| \to min$$



Dimensionality Reduction with SVD

$$A = U\Sigma V^T$$

$$A_k = U_k\Sigma_k V_k^T = (U_k\Sigma_k)V_k^T = U_k(\Sigma_k V_k^T)$$



$$U \in \mathbb{R}^{m\times r}; UU^T = I$$

$$V \in \mathbb{R}^{n\times r}; VV^T = I$$

$$\Sigma = diag(\sigma_1, ..., \sigma_r); r = rank(M)$$

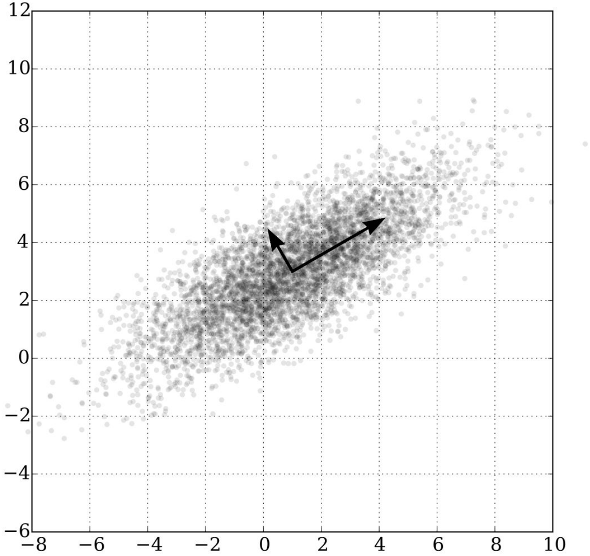
Theorem (Eckart-Young)

- Truncated SVD gives best low-rank approximation for a given matrix A
- More formally,

$$A_k = U_k \Sigma_k V_k^T$$
$$\forall B_k : rank(B_k) = k$$
$$||A - B_k||_F \geq ||A - A_k||_F$$

PCA: Projection into Subspace

- Project all data points into a smaller dimension subspace
- Maximize *variance* along new basis vectors



PCA: Projection into Subspace

$$X = U\Sigma V^T$$

orthogonal

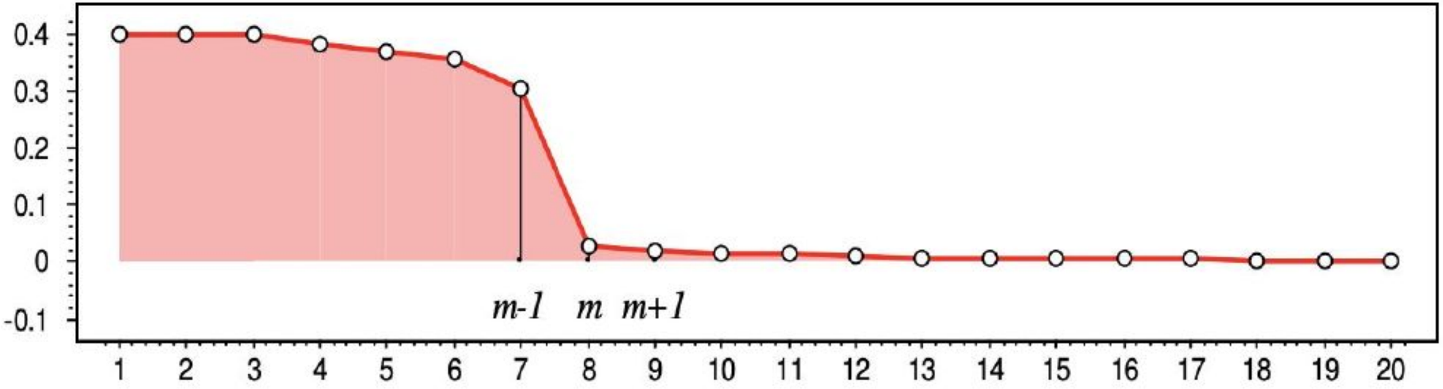
orthogonal

diagonal: sigmas ~ variance

- Consider columns of matrix V new basis vectors: **principal directions**
- Columns of matrix US are called **principal components** of the data
- Singular values are sorted: truncated SVD gives the best projection of dim K

PCA: Effective Dimensionality

- Often data is noisy and has non-informative features
- Get rid of low-variance components in PCA



$$E_m = \frac{\|GU^\intercal - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \cdots + \lambda_n}{\lambda_1 + \cdots + \lambda_n} \leqslant \varepsilon.$$

PCA in Practice

- Above said is correct only if X is centered
    - Normalize data before PCA!
- Dimensionality reduction:

$$X_k = U_k \Sigma_k$$

- Reconstruction:

$$\overline{X} = U_k \Sigma_k V_k^T$$

PCA in Practice: Examples

- Word embeddings visualization

Let's walk through space...

Image source: https://github.com/yandexdataschool/nlp_course/tree/2019/week01_embeddings

PCA in Practice: Examples

- Eigenfaces: image examples

Image source: https://github.com/yandexdataschool/nlp_course/tree/2019/week01_embeddings

PCA in Practice: Examples

- Eigenfaces: top-16 components

Image source: https://github.com/yandexdataschool/nlp_course/tree/2019/week01_embeddings

PCA in Practice: Examples

- Eigenfaces: reconstruction with n=50

Image source: https://github.com/yandexdataschool/nlp_course/tree/2019/week01_embeddings

Revise

- Multiclass aggregation strategies
  - One vs Rest
  - One vs One
- Metrics in classification (again):
  - Precision, Recall, F-score
  - ROC curve, PR curve, AUC
  - Confusion matrix
- Dimensionality reduction and PCA
  - Connections with SVD

Next time

- Train-validation-test split
- Validation Strategies
- Bias-variance Trade-off

Thanks for attention!

Questions?

girafe ai