Machine Learning
Linear Classification
Logistic Regression

Young && Yandex ШАД

girafe ai

Radoslav Neychev

Recap

Lecture 2: Linear Regression

01 Linear Models overview

02 Regression problem statement

03 Linear Regression analytical solution:

- Gauss-Markov theorem (BLUE)
- Instability

04 Regularization:

- L2 aka Ridge: Analytical solution
- L1 aka LASSO: Weights decay rule
- Elastic Net

05 Metrics in regression

06 Model building cycle

- Train
- Validation
- Test

Outline

01 Linear classification

- Margin
- Loss functions

02 Logistic regression

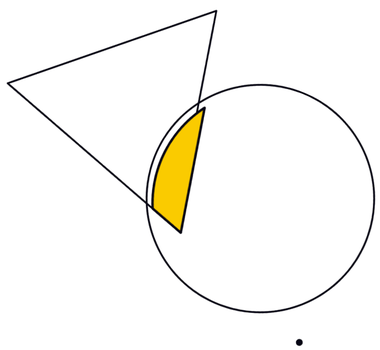
- Sigmoid derivation
- Maximum Likelihood Estimation
- Logistic loss
- Probability calibration

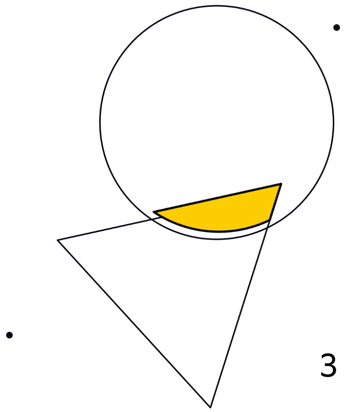
03 Multiclass aggregation strategies

- One vs Rest
- One vs One

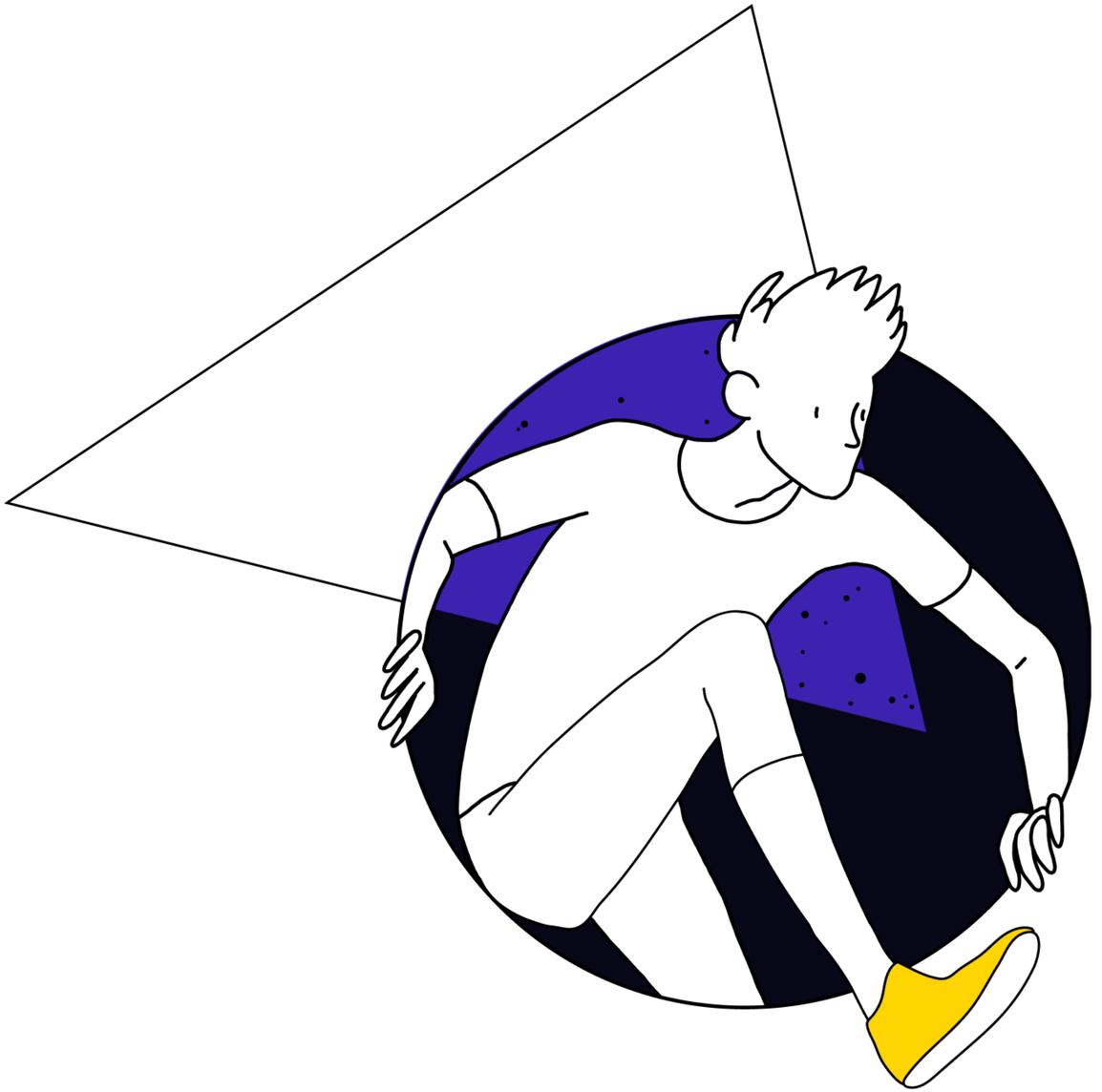
04 Metrics in classification

- Accuracy, Balanced accuracy
- Precision, Recall, F-score
- ROC curve, PR curve, AUC
- Confusion matrix





Linear Classification



01

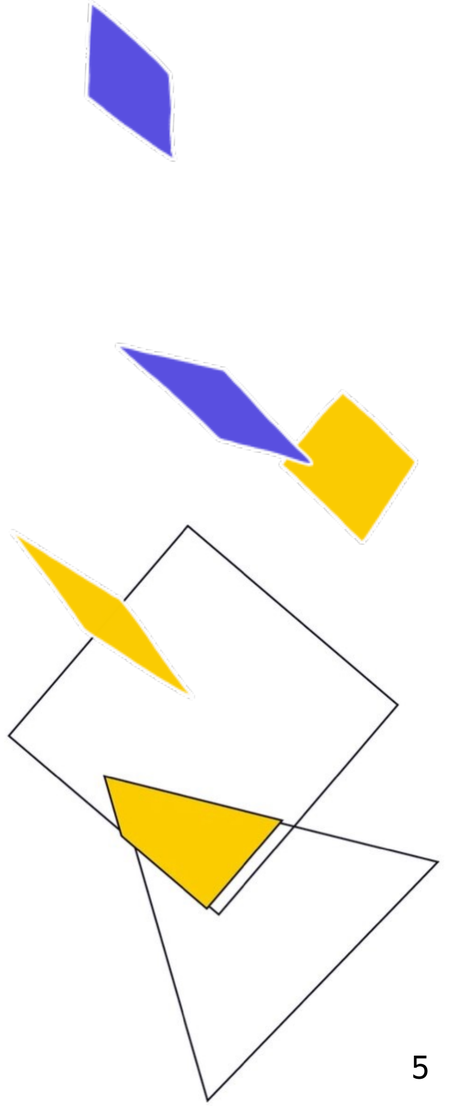
Classification problem

Let’s denote:

- Training set $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$,where
  - $\boldsymbol{x}^{(i)} \in \mathbb{R}^p, y^{(i)} \in \{C_1, \ldots, C_K\}$ for classification
- Model $c(\boldsymbol{x})$ predicts class label or classes probability vector for every object
- Loss function $L(\boldsymbol{x}, y, f)$ that should be minimized

Consider binary classification for now:

$$y^{(i)} \in \{+1, -1\}$$



Linear classifier

The most simple linear classifier:

$$c(\boldsymbol{x}) = \begin{cases} 1 & \text{if } f(\boldsymbol{x}) \geq 0 \\ -1 & \text{if } f(\boldsymbol{x}) < 0 \end{cases}$$

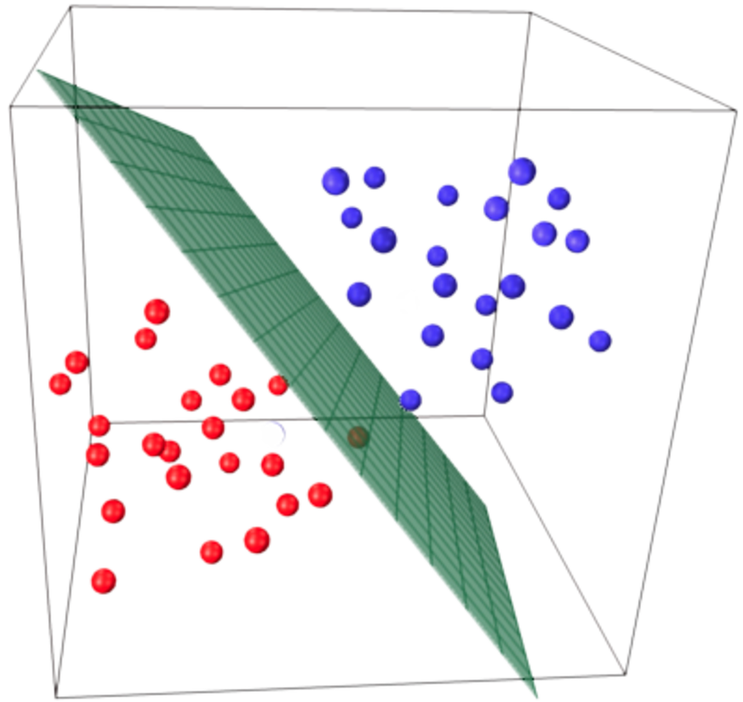
or equivalently:

$$c(\boldsymbol{x}) = \text{sign}\big(f(\boldsymbol{x})\big) = \text{sign}(\boldsymbol{x}^\top \boldsymbol{w})$$

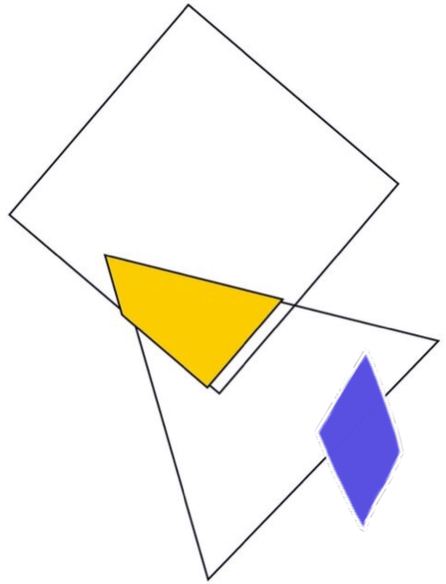
for linear model

Why cutoff value is fixed?
(bias term is implied)

Geometrical interpretation:
hyperplane dividing space into two subspaces



Margin



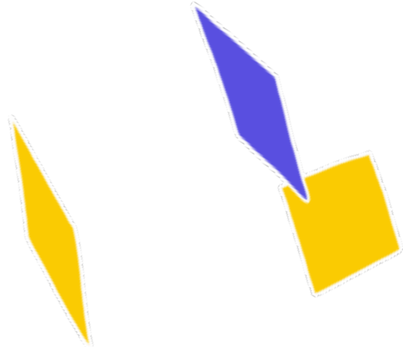
Let’s define linear model’s Margin as

$$M^{(i)} = y^{(i)} \cdot f(\boldsymbol{x}^{(i)}) = y^{(i)} \cdot \boldsymbol{x}^{(i)\top}\boldsymbol{w}$$

Main property:
negative margin reveals misclassification

$$M^{(i)} \geq 0 \Leftrightarrow y^{(i)} = c(\boldsymbol{x}^{(i)})$$

$$M^{(i)} < 0 \Leftrightarrow y^{(i)} \neq c(\boldsymbol{x}^{(i)})$$



Weights choice

Remembering old paradigm

$$\text{Empirical risk} = \sum_{\text{by objects}} \text{Loss on object} \to \min_{\text{model params}}$$

Disadvantages:

- Not differentiable
- Overlooks confidence

Essential loss is misclassification

$$L_{\text{miss}}(y^{(i)}, \hat{y}^{(i)}) = \left[y^{(i)} \neq \hat{y}^{(i)}\right] = \left[M^{(i)} \leq 0\right]$$

Solution:

Estimate it with a smooth function

Iverson bracket

$$[P] = \begin{cases} 1, & \text{if } P \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

Square loss

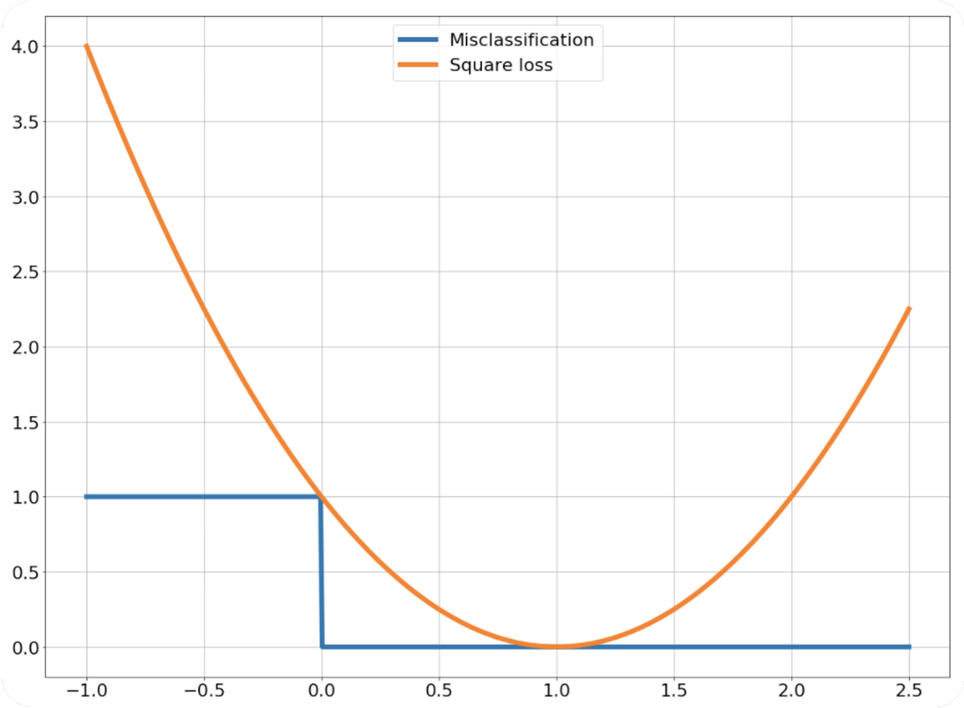
Let’s treat classification problem as regression problem:

$$y^{(i)} \in \{+1, -1\} \subset \mathbb{R}$$

thus we optimize MSE

$$L_{\text{MSE}} = (y^{(i)} - \boldsymbol{x}^{(i)\top}\boldsymbol{w})^2 = \frac{\left((y^{(i)})^2 - y^{(i)} \cdot \boldsymbol{x}^{(i)\top}\boldsymbol{w}\right)}{(y^{(i)})^2}$$

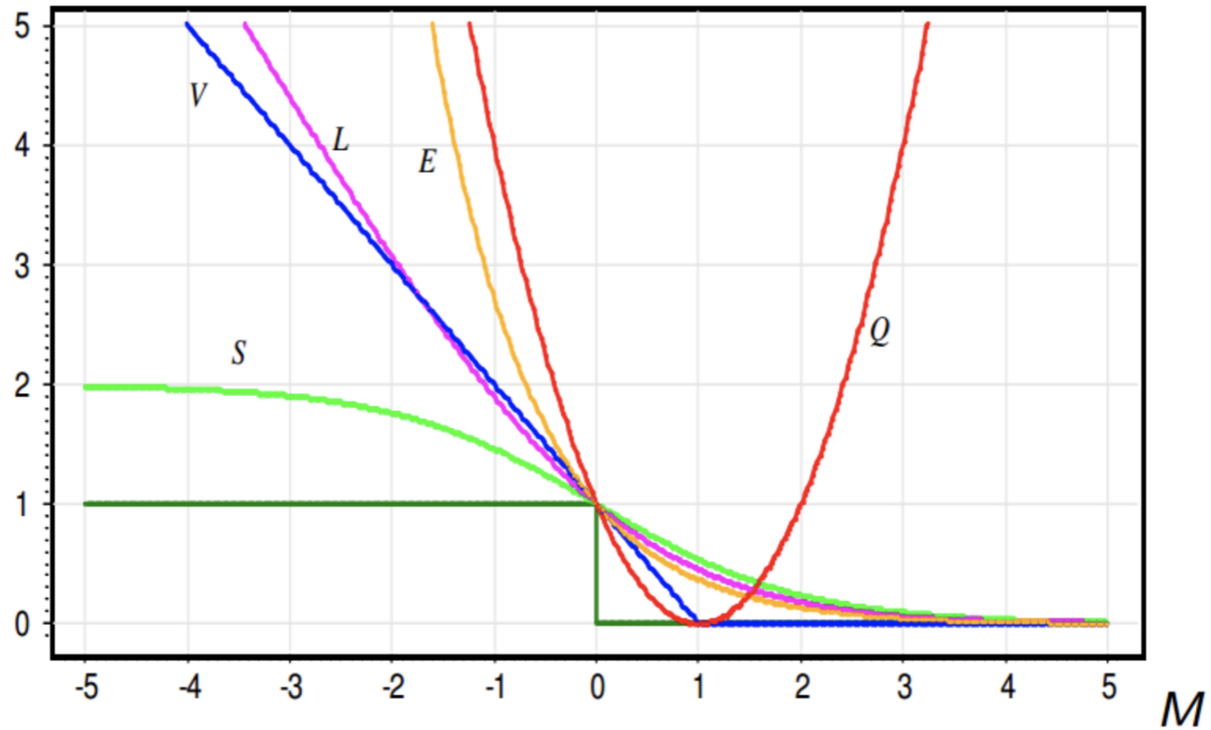
$$= (1 - y^{(i)}\boldsymbol{x}^{(i)\top}\boldsymbol{w})^2 = (1 - M^{(i)})^2$$



Advantage: already solved

Disadvantage: penalizes for high confidence

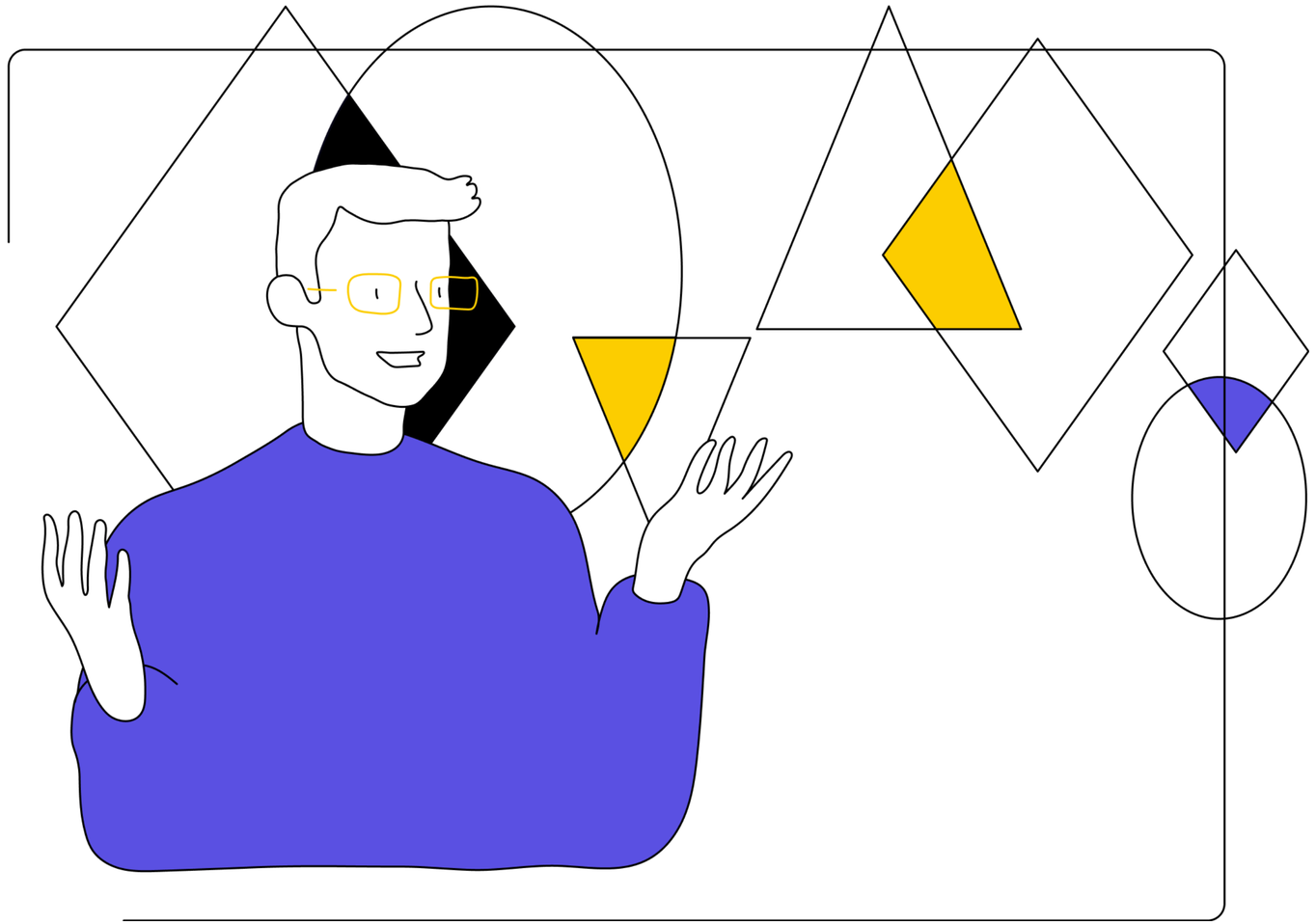
Other losses



square loss	$Q(M) = (1 - M)^2$
hinge loss	$V(M) = (1 - M)_+$
savage loss	$S(M) = 2(1 + e^M)^{-1}$
logistic loss	$L(M) = \log(1 + e^{-M})$
exponential loss	$E(M) = e^{-M}$

Loss functions for classification

Logistic Regression



02

Intuition

01 Let’s try to predict probability of an object to have positive class

$$p_+ = P(y = 1|\boldsymbol{x}) \in [0, 1]$$

02 But all we can predict is a real number!

$$\boldsymbol{x}^\top \boldsymbol{w} \in \mathbb{R}$$

03 Time for some tricks

$$\log\left(\frac{p_+}{1 - p_+}\right) \in \mathbb{R}$$

04 Reverse to closed form

$$\frac{p_+}{1 - p_+} = \exp(\boldsymbol{x}^\top \boldsymbol{w}) \in \mathbb{R}^+$$

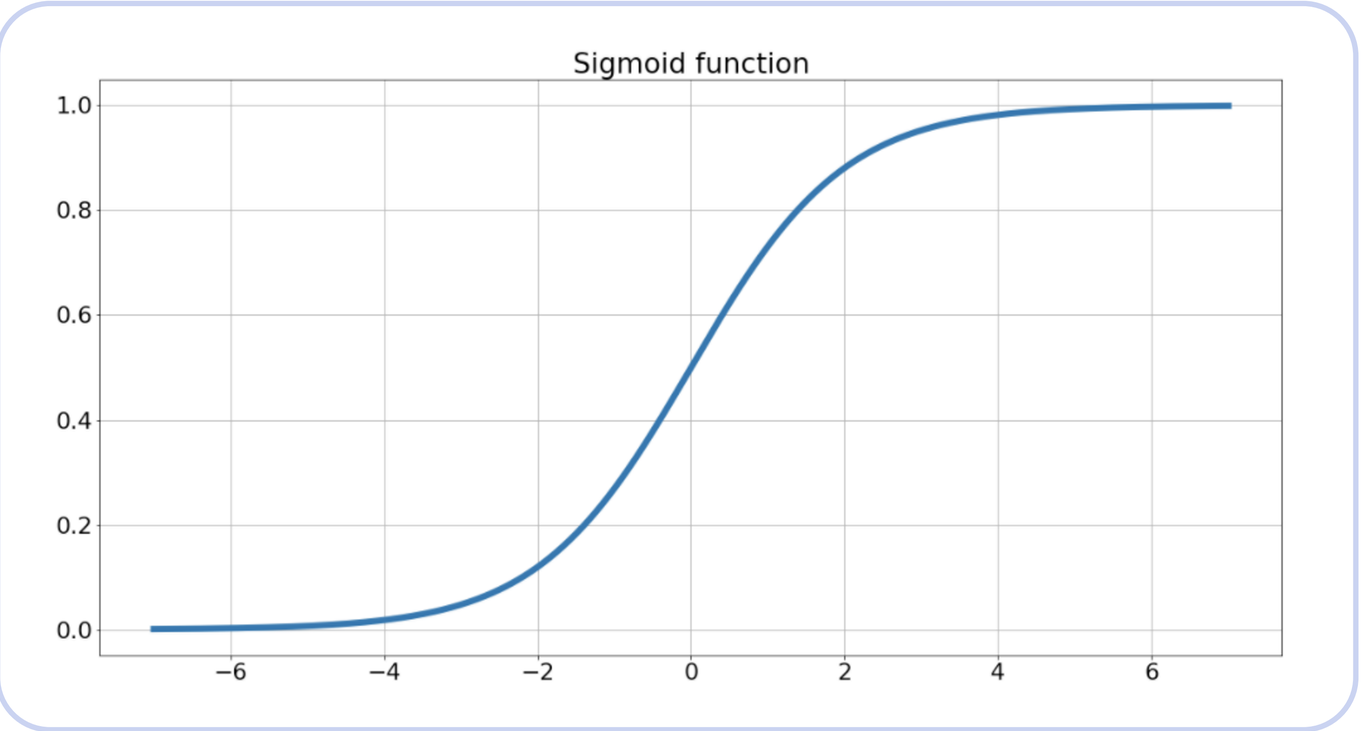
05 Here is the match

$$p_+ = \frac{1}{1 + \exp(-\boldsymbol{x}^\top \boldsymbol{w})} = \sigma(\boldsymbol{x}^\top \boldsymbol{w})$$

Sigmoid (aka logistic) function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Sigmoid is odd relative to (0, 0.5) point



Symmetric property:

$$1 - \sigma(z) = \sigma(-z)$$

Derivative:

$$\sigma(z)' = \sigma(z)\big(1 - \sigma(z)\big)$$

Maximum Likelihood Estimation

Just to remind

$$\log \mathcal{L}(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \log \sigma(M^{(i)}) = -\sum_{i=1}^{n} \log(1 + \exp\left(-M^{(i)}\right))$$

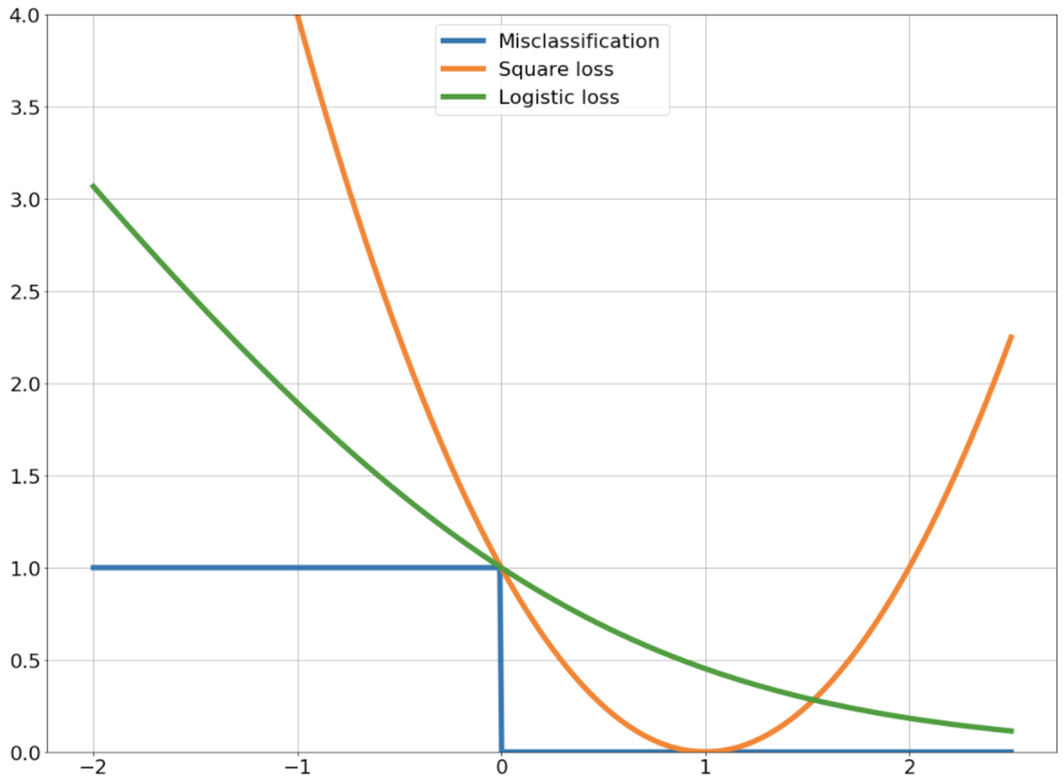
Calculating probabilities for objects

$$\text{if } y = +1 : P(\boldsymbol{x}, y|\boldsymbol{w}) = \sigma(\boldsymbol{x}^\top \boldsymbol{w}) = \sigma(M)$$

$$\text{if } y = -1 : P(\boldsymbol{x}, y|\boldsymbol{w}) = 1 - \sigma(\boldsymbol{x}^\top \boldsymbol{w}) = \sigma(M)$$

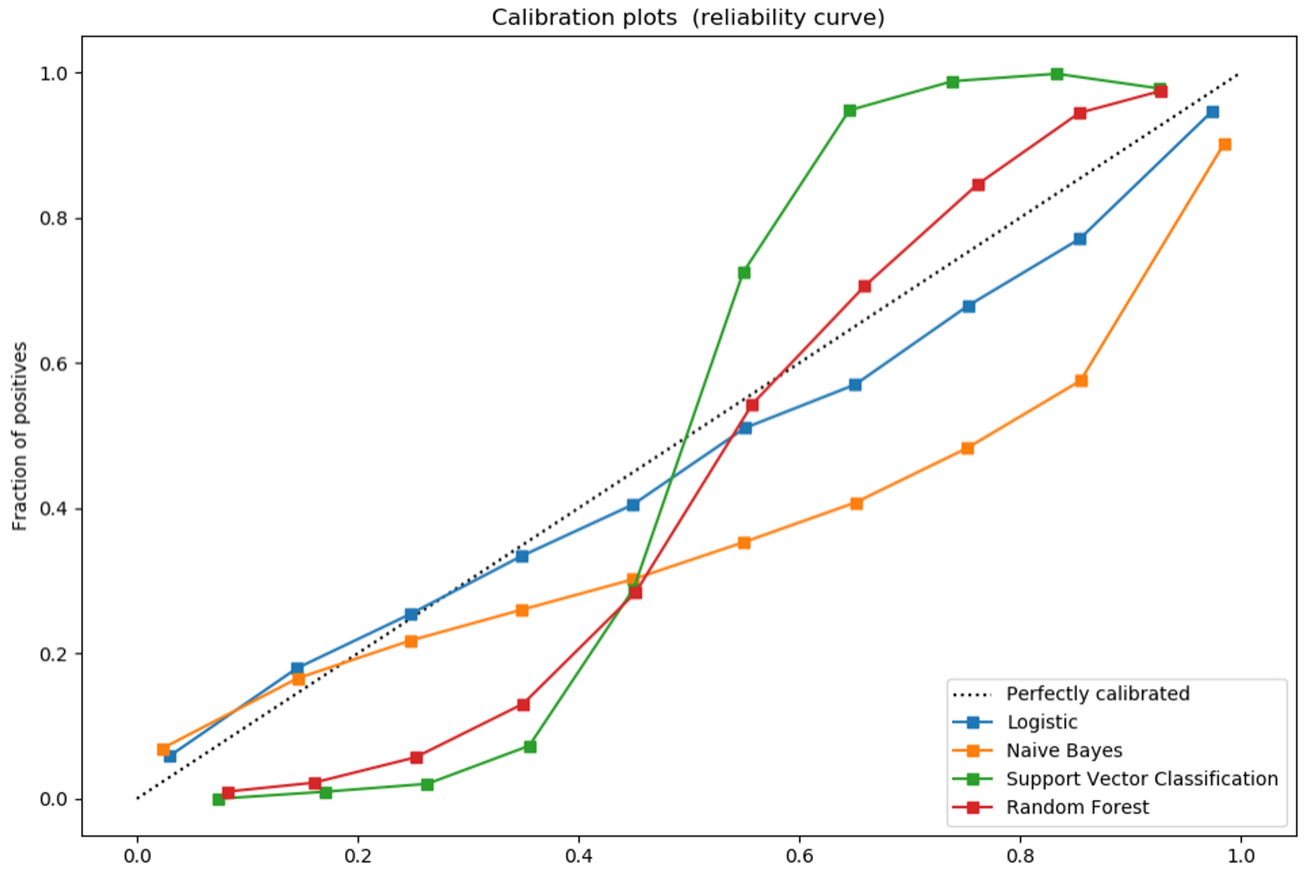
$$\log \mathcal{L}(\boldsymbol{w}|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \log \sigma(M^{(i)}) = -\sum_{i=1}^{n} \log\left(1 + \exp(-M^{(i)})\right)$$

Logistic loss



$$L_{\text{Logistic}} = \log\left(1 + \exp(-M)\right)$$

Probability calibration



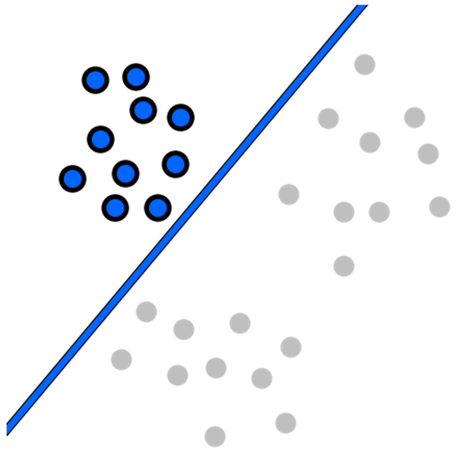
By using Logistic Regression we generate a Bernoulli distribution in each point of space

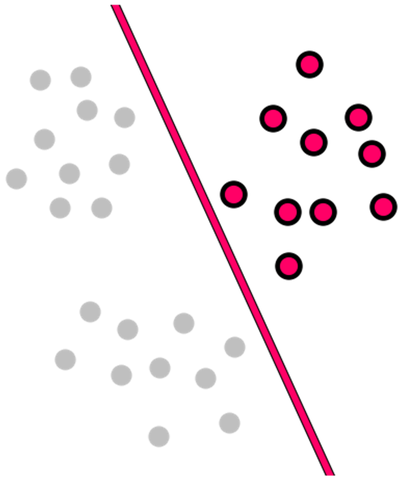
Calibration discussion

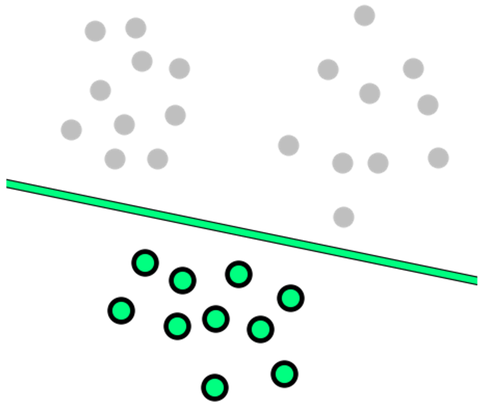
Multiclass aggregation strategies

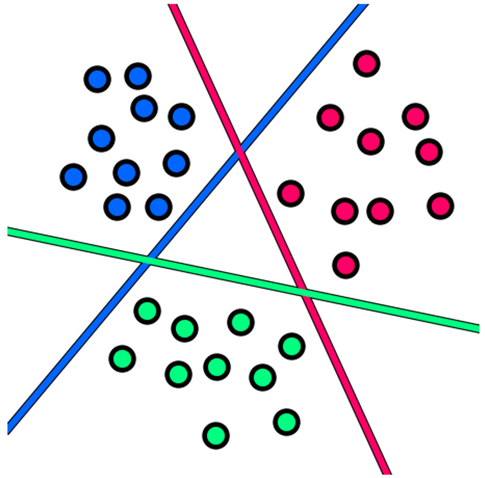
03

One vs Rest



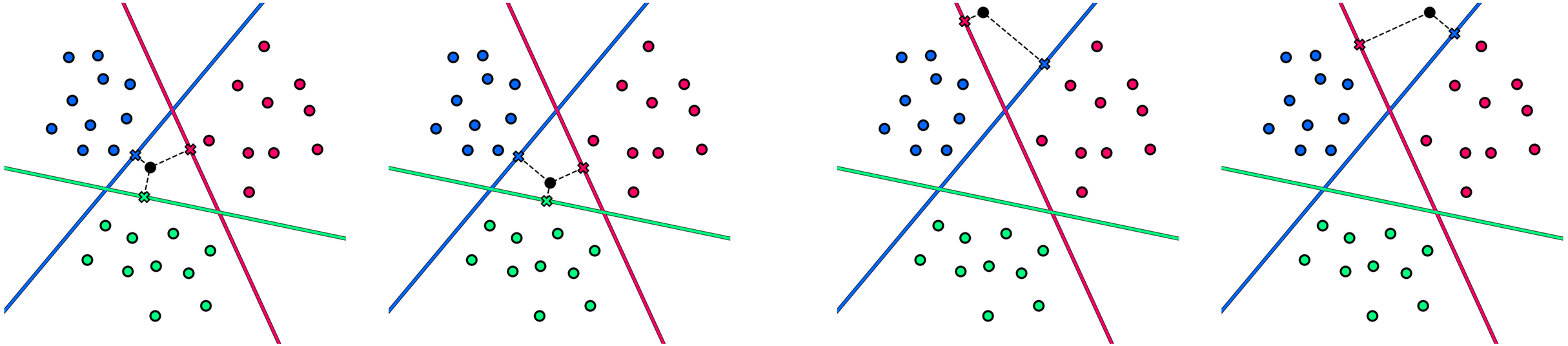




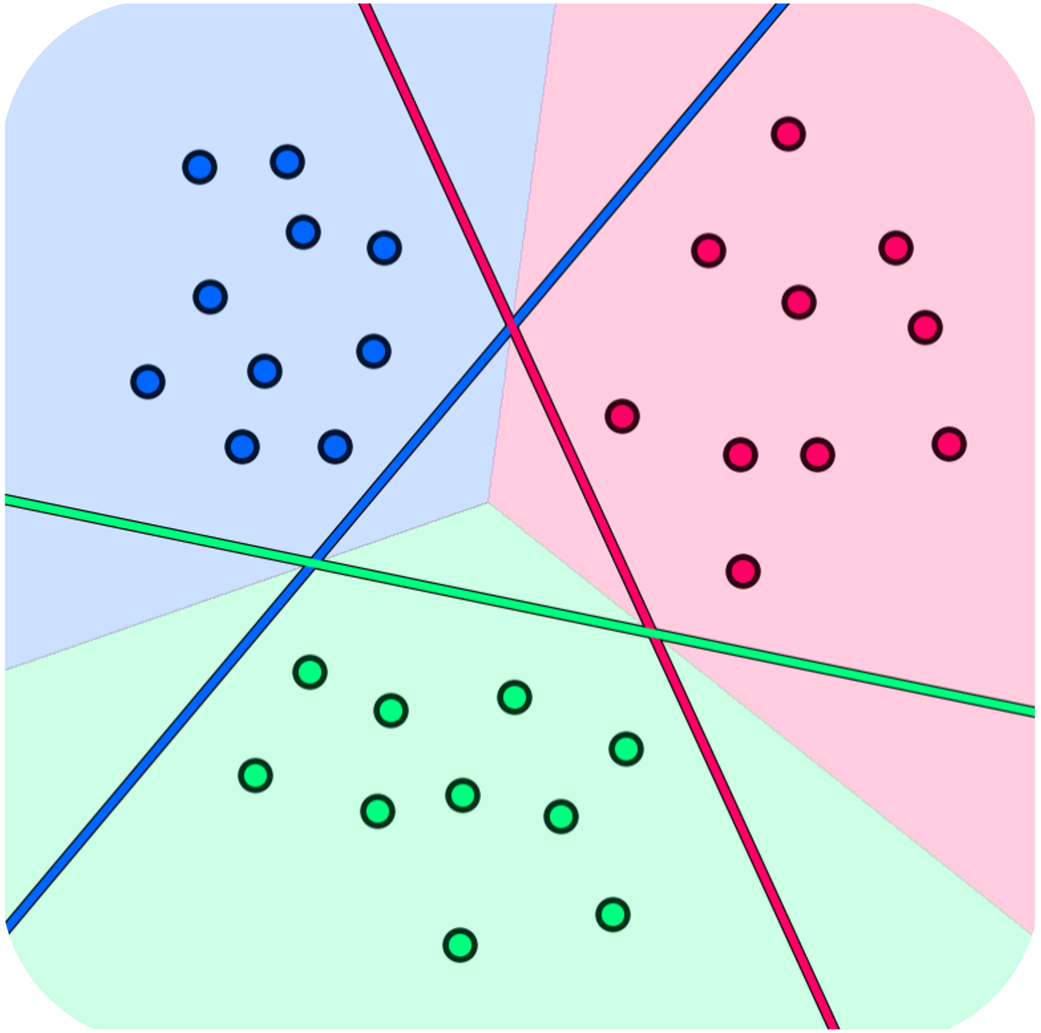


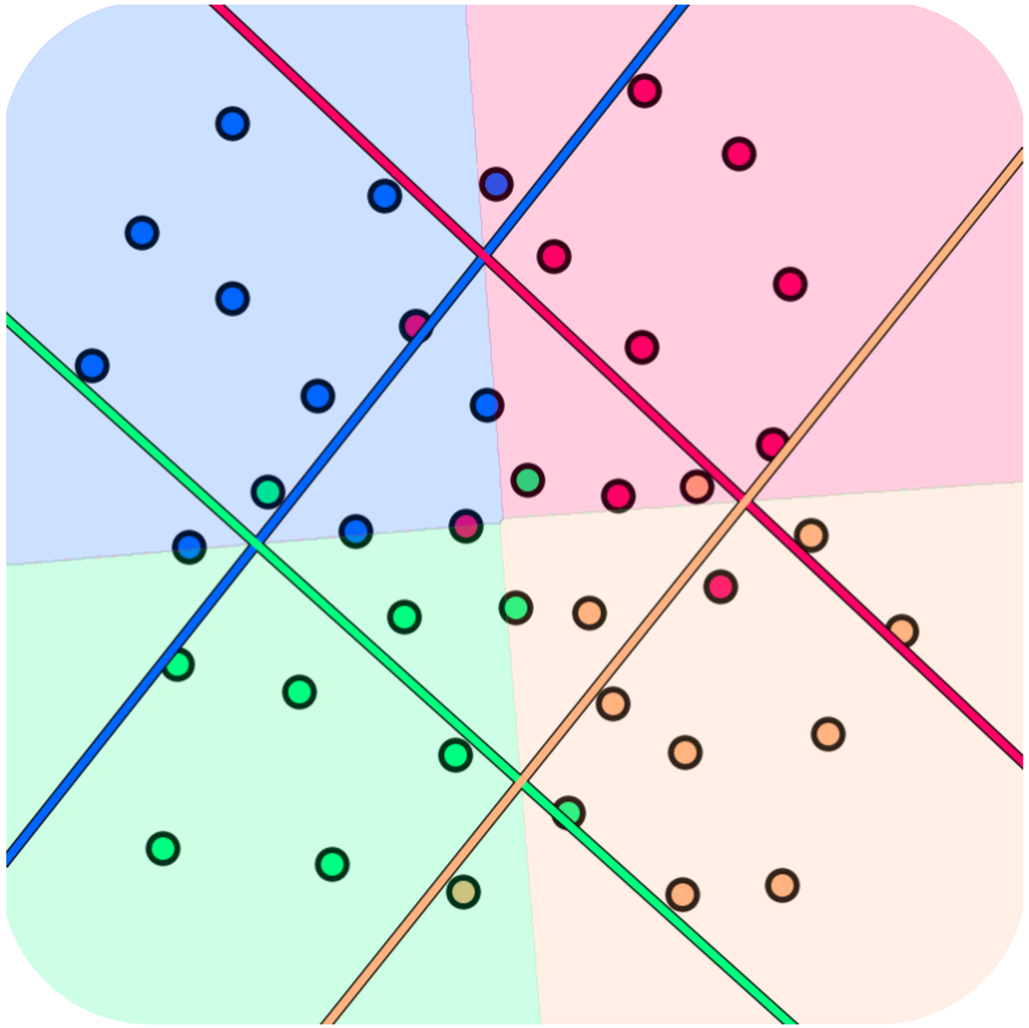
Images source

One vs Rest: unclassified regions

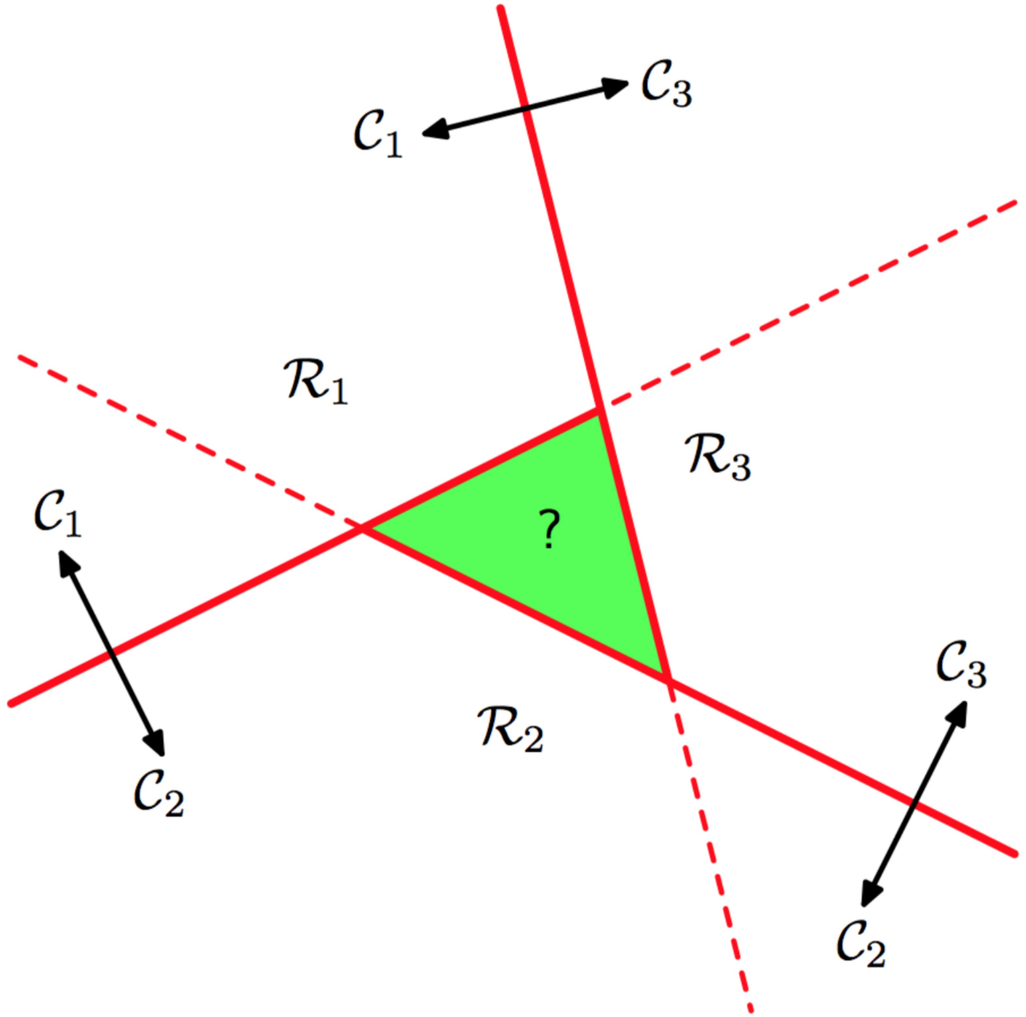


One vs Rest: final result

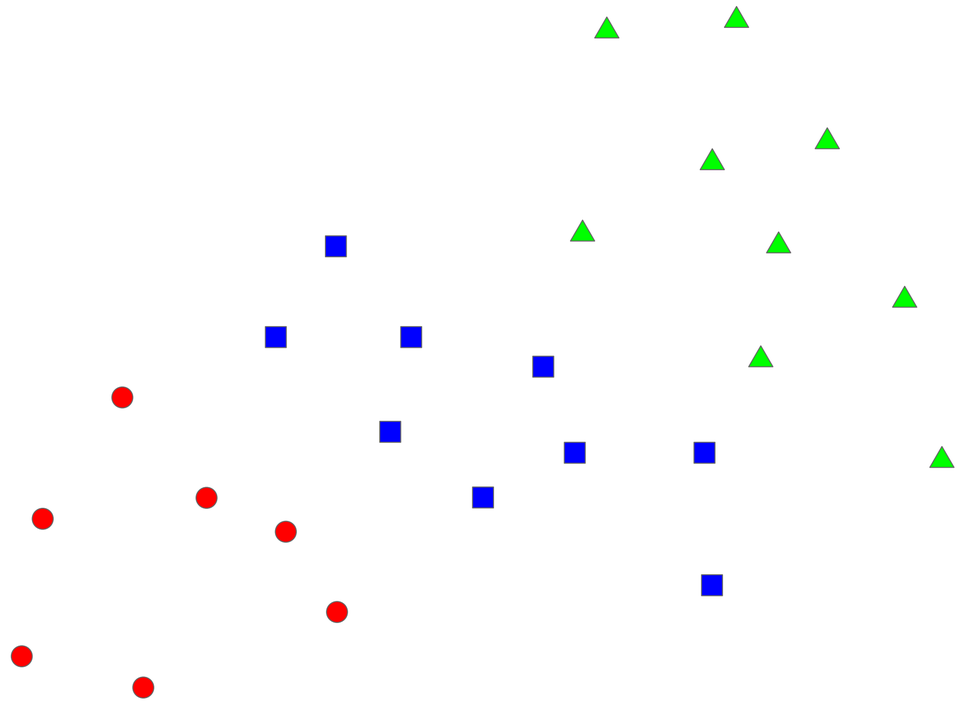




One vs One

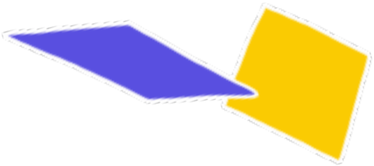


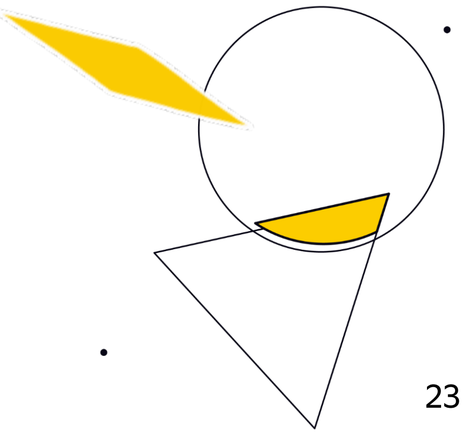
Failure case?



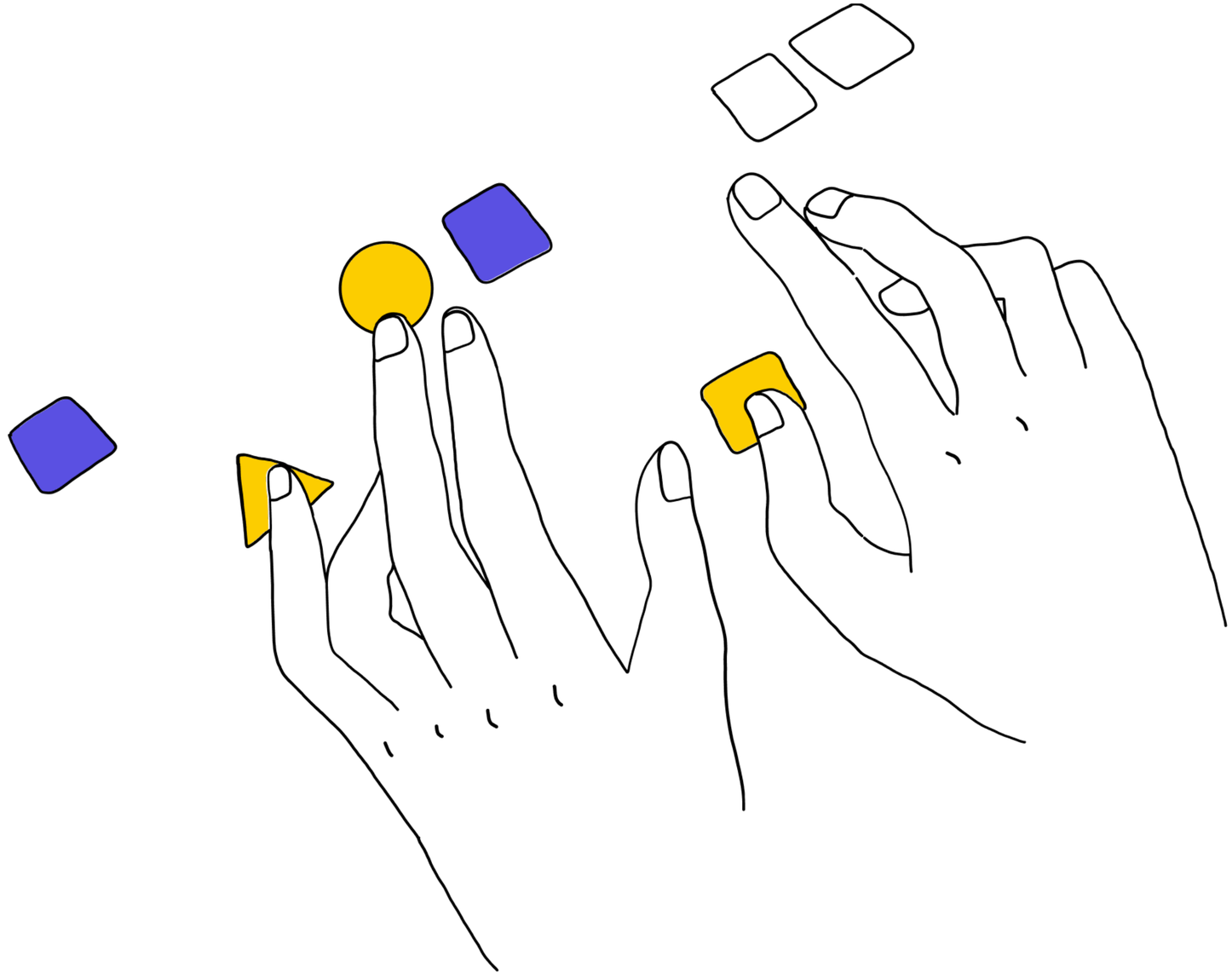
Summary

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





Metrics in classification



04

Metrics

01 **Accuracy**
Balanced accuracy

02 Precision

03 Recall

04 F-score

05 **ROC curve**
ROC-AUC

06 **PR curve**
PR-AUC

07 Multiclass generalizations

08 Confusion matrix

Accuracy

Number of right classifications

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}[y^{(i)} = \hat{y}^{(i)}]$$

target: 1 0 1 0 0 0 0 1 0 0

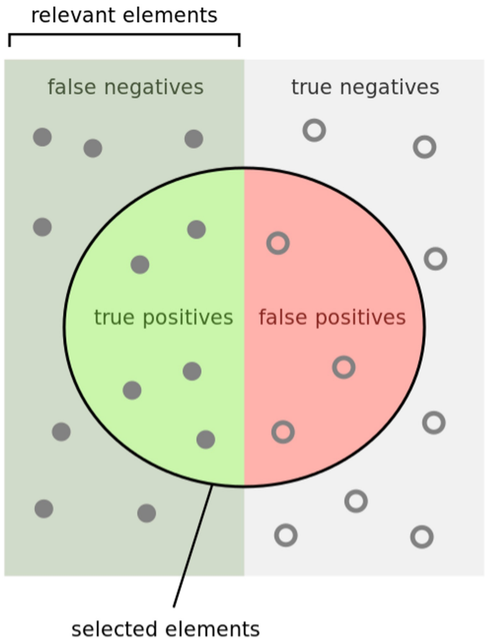
predicted: 0 0 1 0 0 0 0 1 1 0

accuracy = 8/10 = 0.8

$$\text{Balanced accuracy} = \frac{1}{C}\sum_{k=1}^{C}\frac{\sum_i[y^{(i)} = k \text{ and } \hat{y}^{(i)} = y^{(i)}]}{\sum_i[y^{(i)} = k]}$$

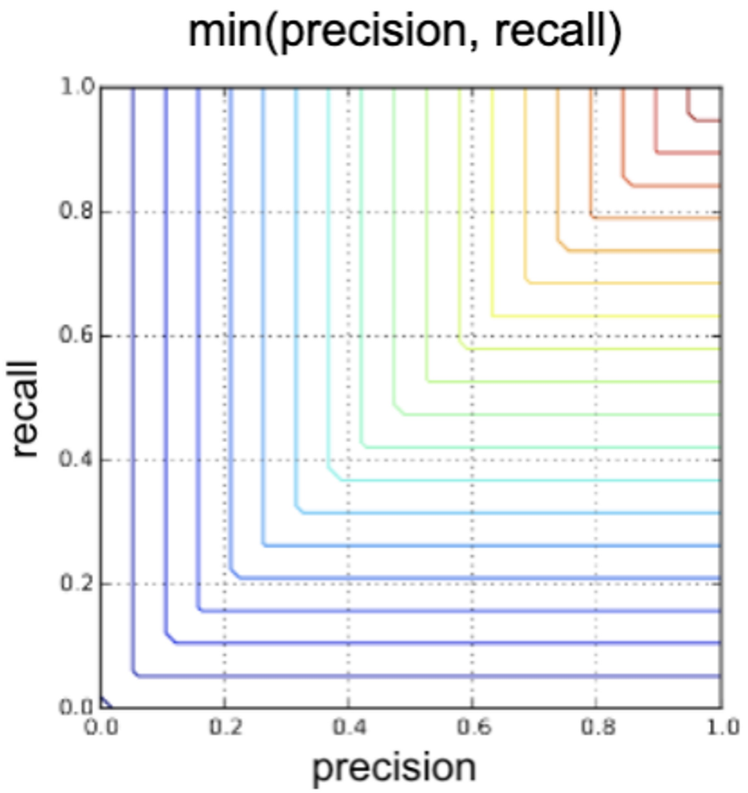
Precision and Recall

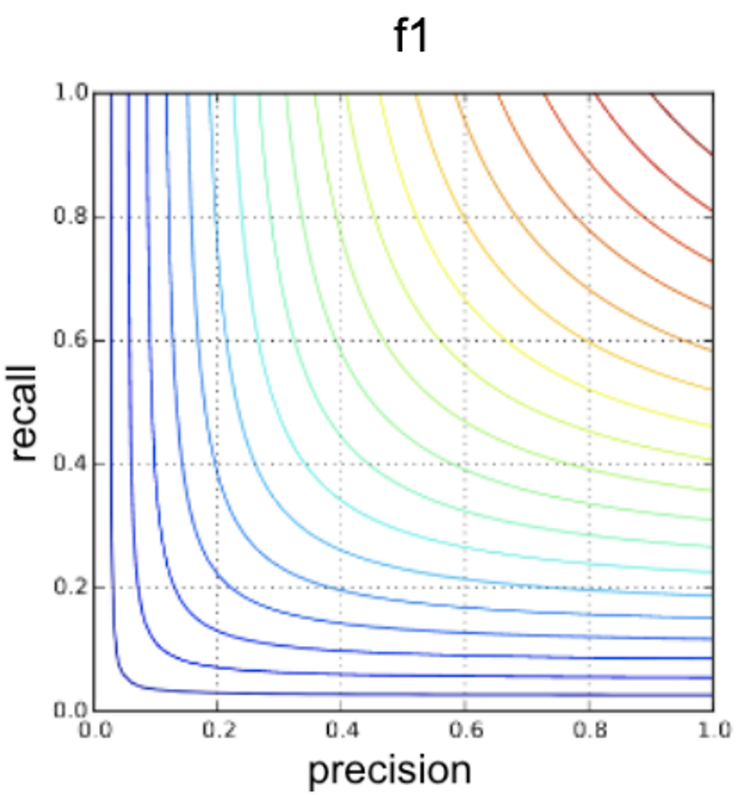
		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**



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

F-score motivation





F-score

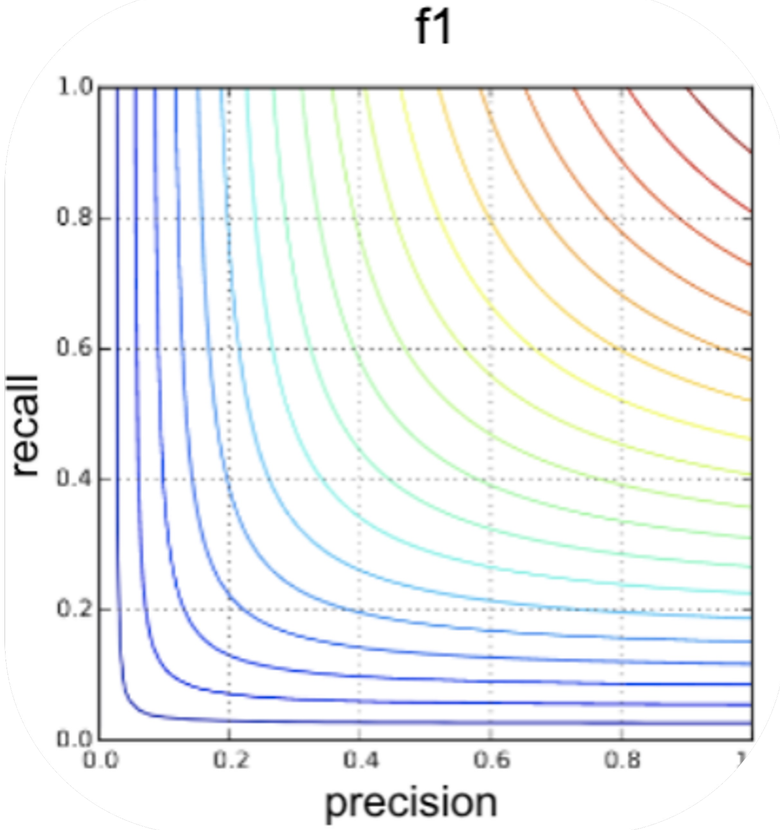
Harmonic mean of precision and recall

Closer to smaller one

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

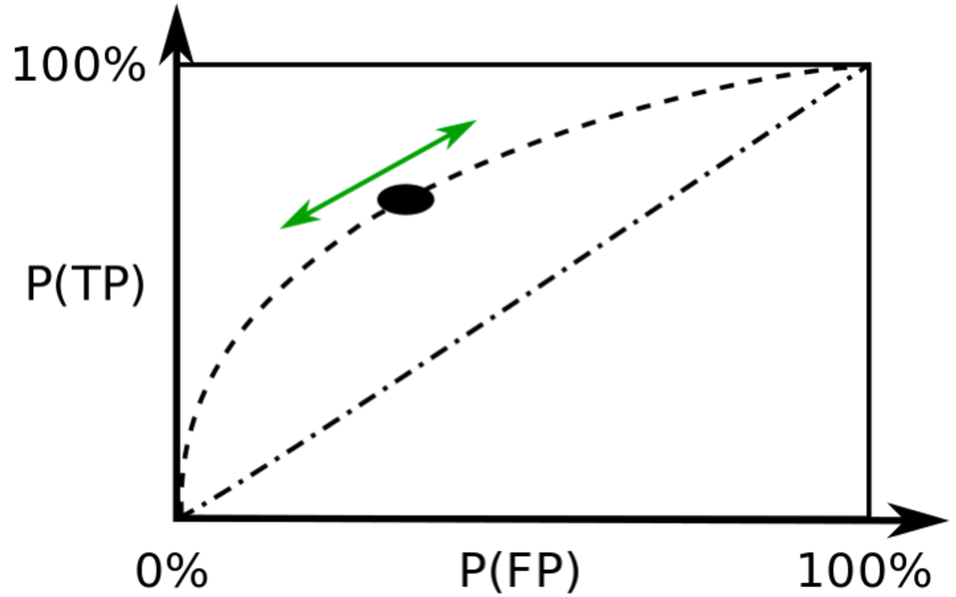
Generalization to different ratio between Precision and Recall

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



Receiver Operating Characteristic (ROC)

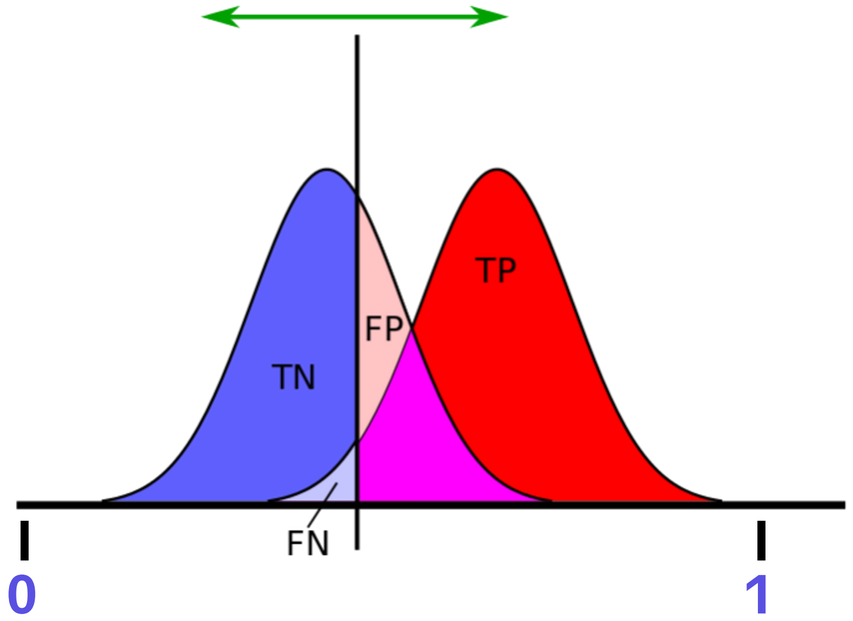
		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**



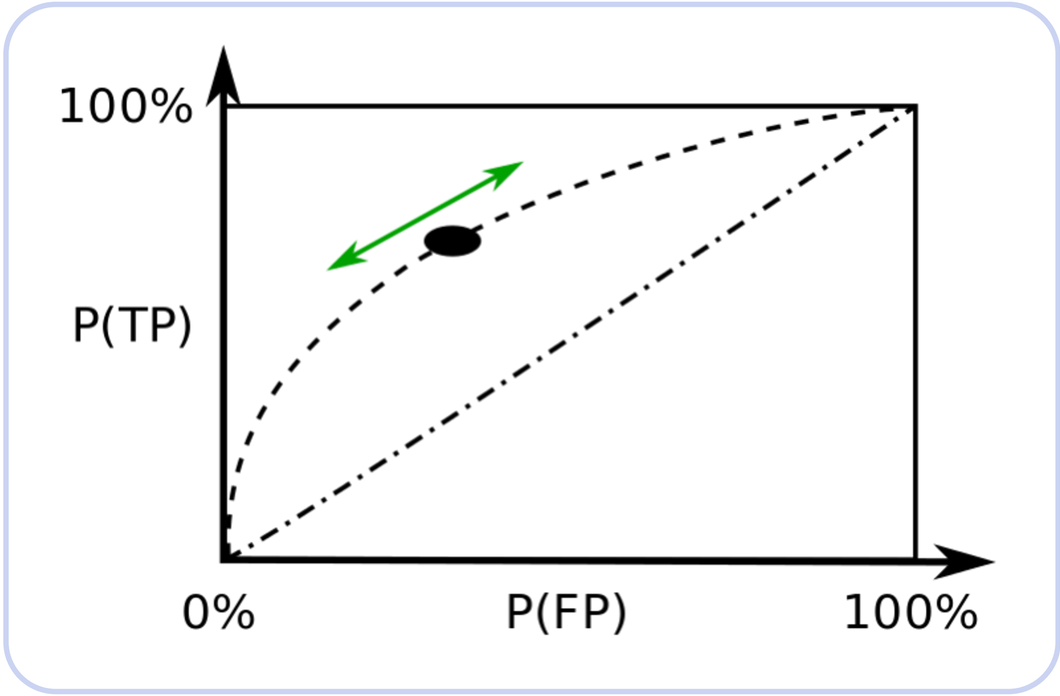
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = \mathrm{Recall}_{\mathrm{negative}} \quad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \mathrm{Recall}_{\mathrm{positive}}$$

Receiver Operating Characteristic (ROC)

01 Classifier needs to predict probabilities



02 Objects get sorted by positive probability



03 Line is plotted as threshold moves

Receiver Operating Characteristic (ROC)

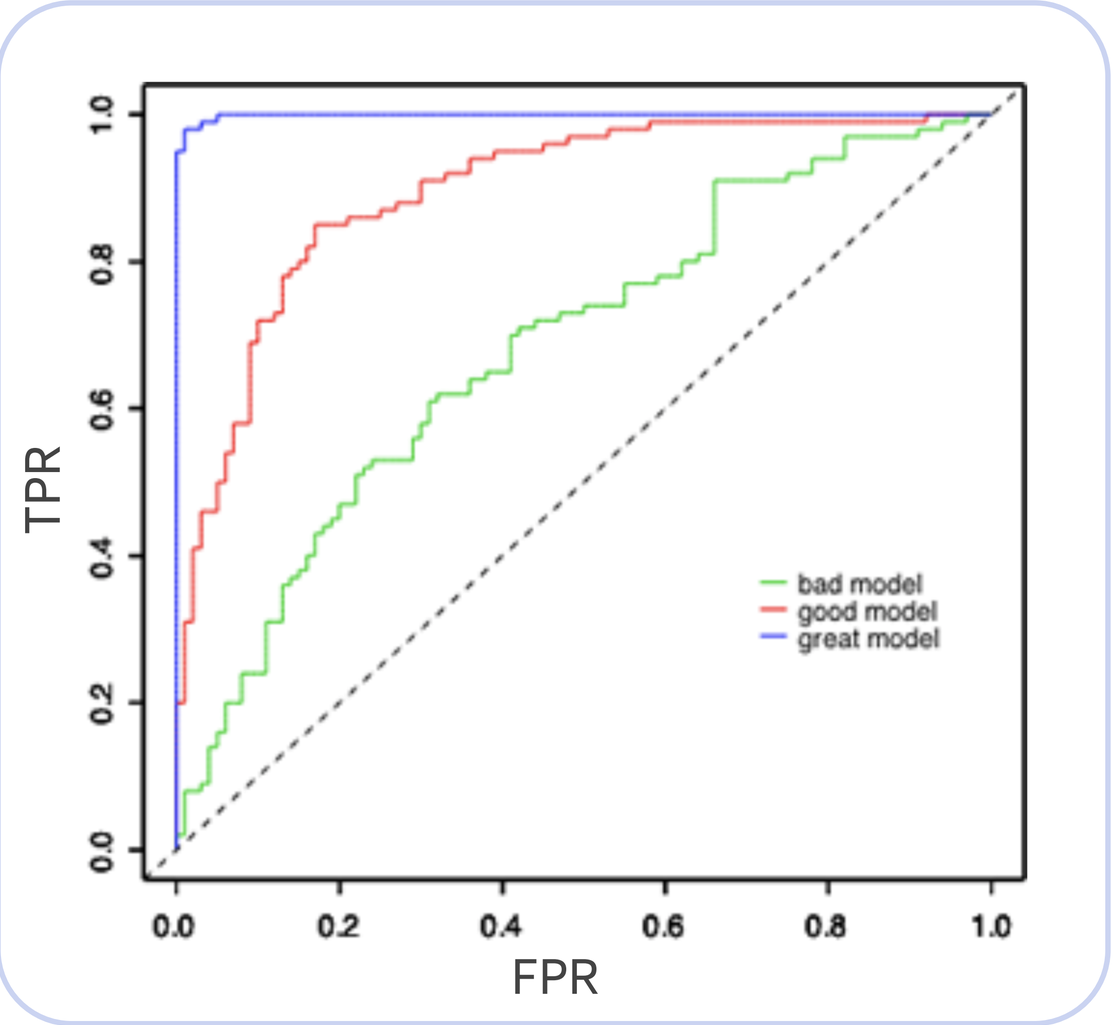
01 Baseline is random predictions

02 Always above diagonal (for reasonable classifier)

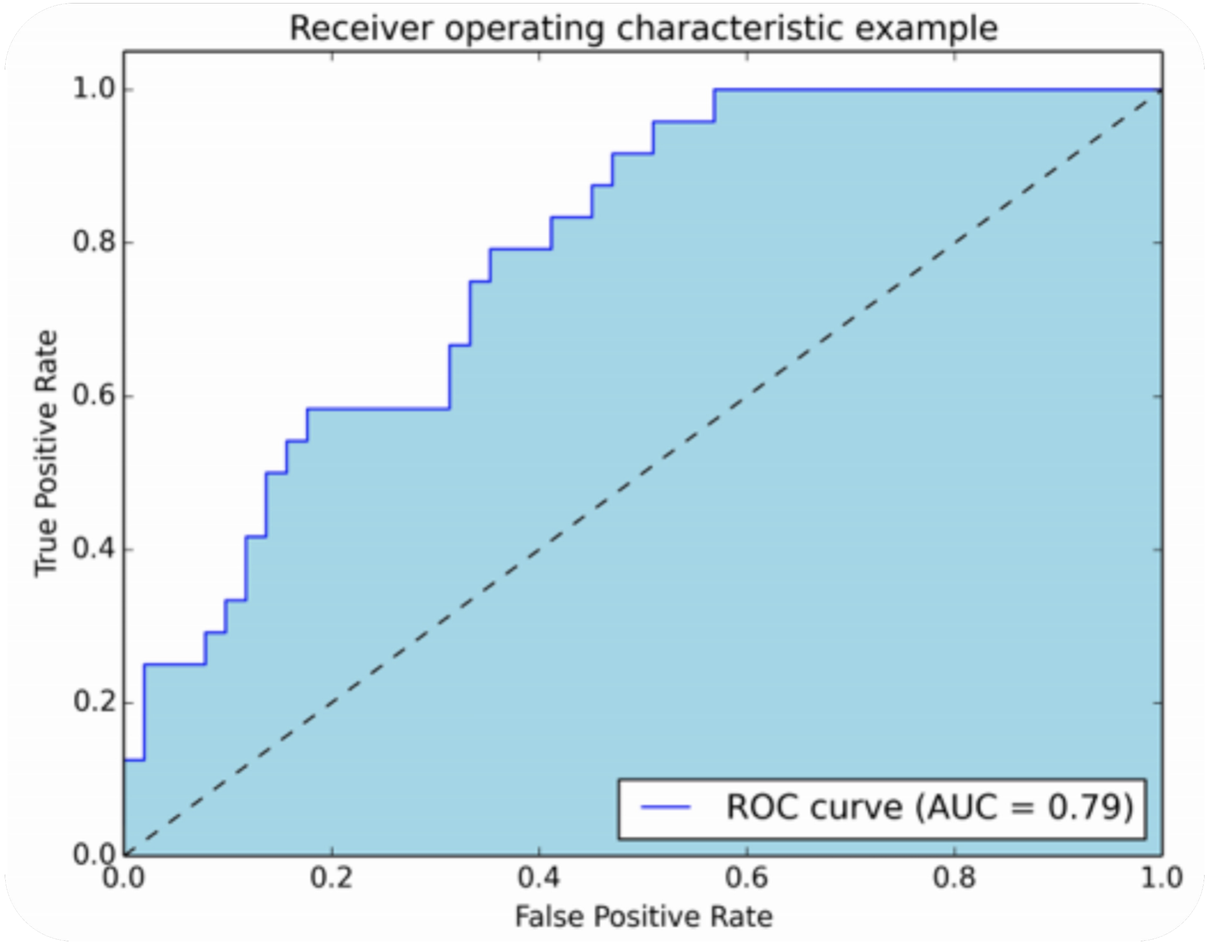
03 If below — change sign of predictions

04 Strictly higher curve means better classifier

05 Number of steps (thresholds) not bigger than dataset



ROC Area Under Curve (ROC-AUC)

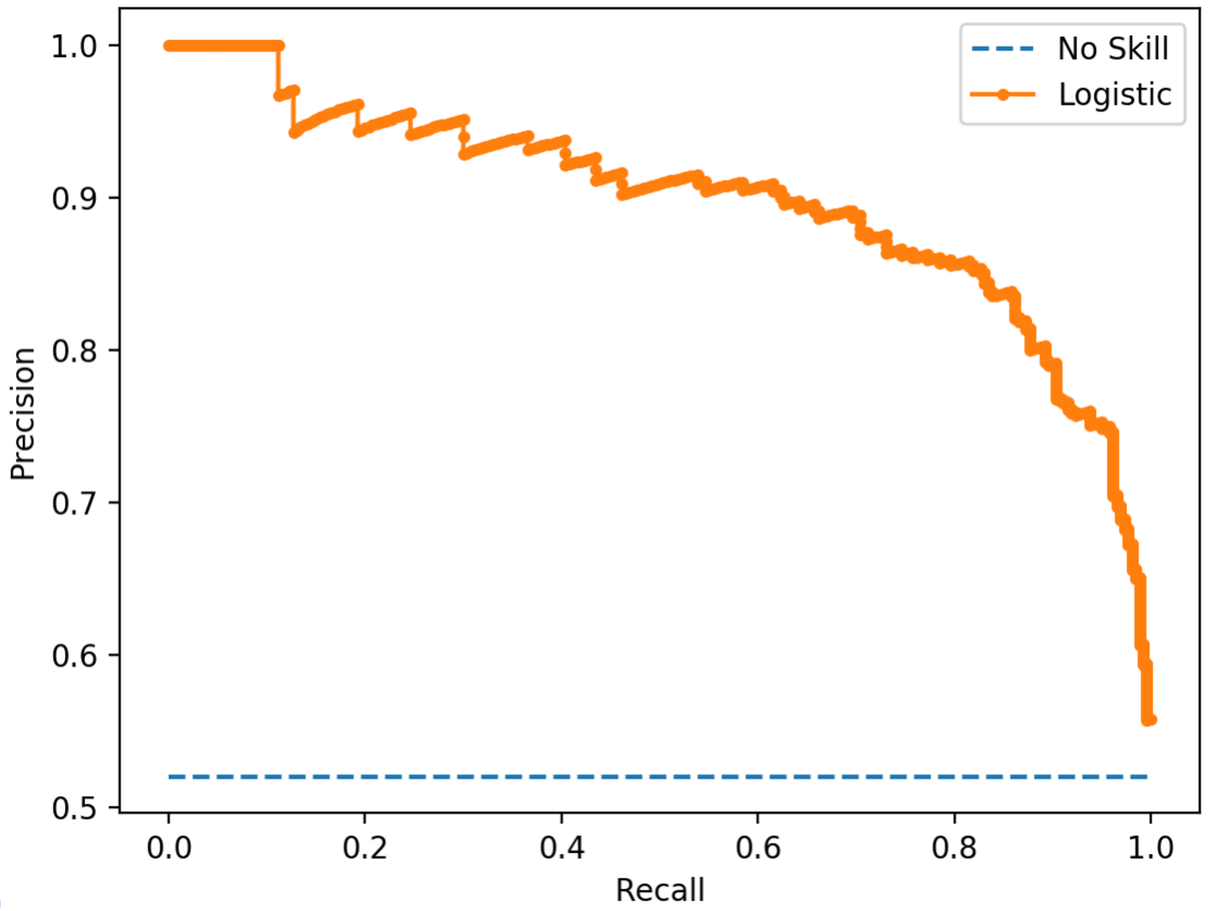


01 Effectively lays in (0.5, 1)

02 Bigger ROC-AUC doesn't imply higher curve everywhere

More explanations with pictures

Precision-Recall Curve



01 AUC is in (0, 1)

02 Source of AP metric
(important for next semester)

Nice article

Multiclass metrics

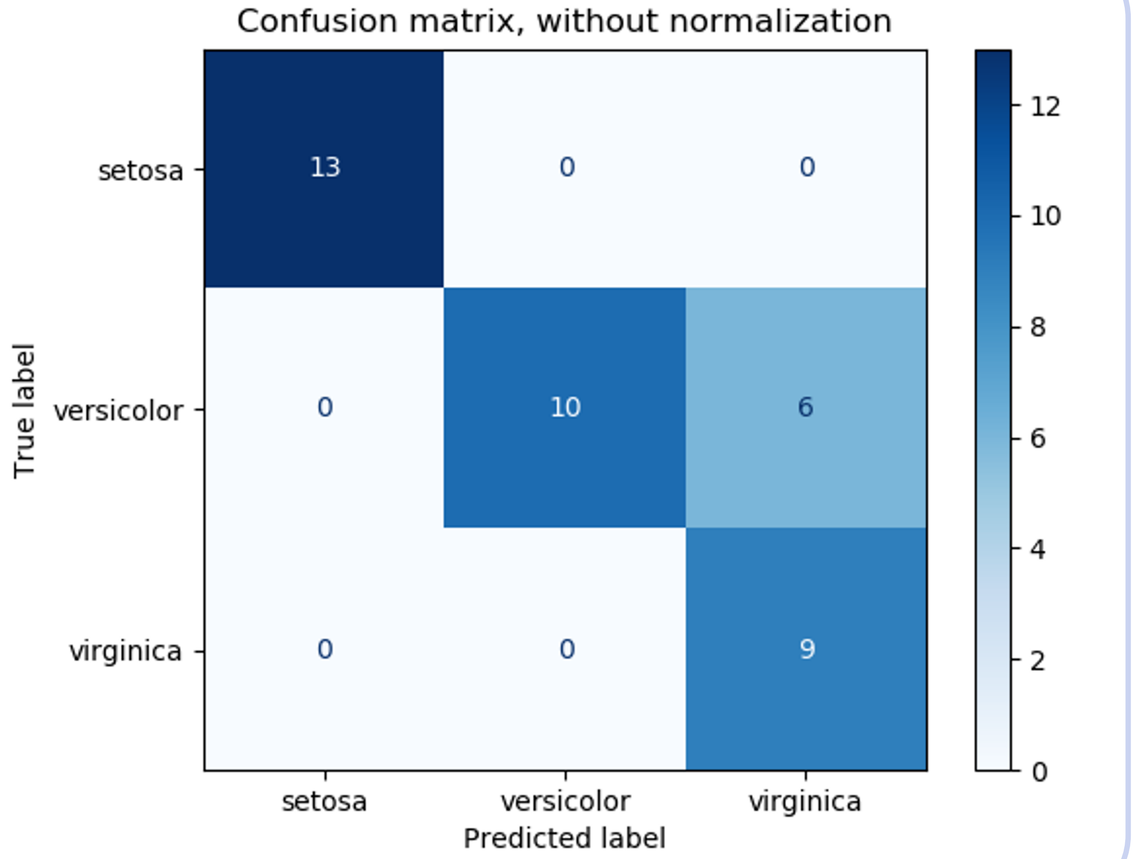
As with linear models we need some magic to measure multiclass problems

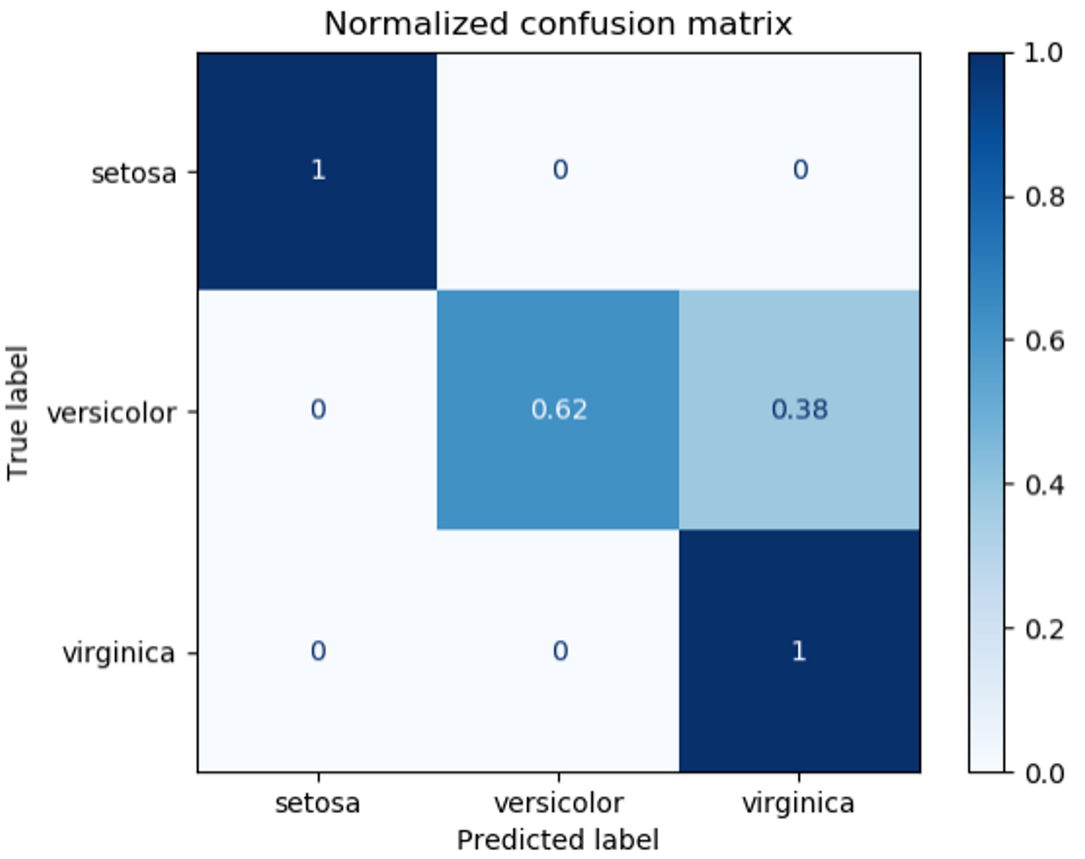
Basically it’s mean of one or another kind

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

Detailed info here here

Confusion matrix





Revise

01 Linear classification

- Margin
- Loss functions

02 Logistic regression

- Sigmoid derivation
- Maximum Likelihood Estimation
- Logistic loss
- Probability calibration

03 Multiclass aggregation strategies

- One vs Rest
- One vs One

04 Metrics in classification

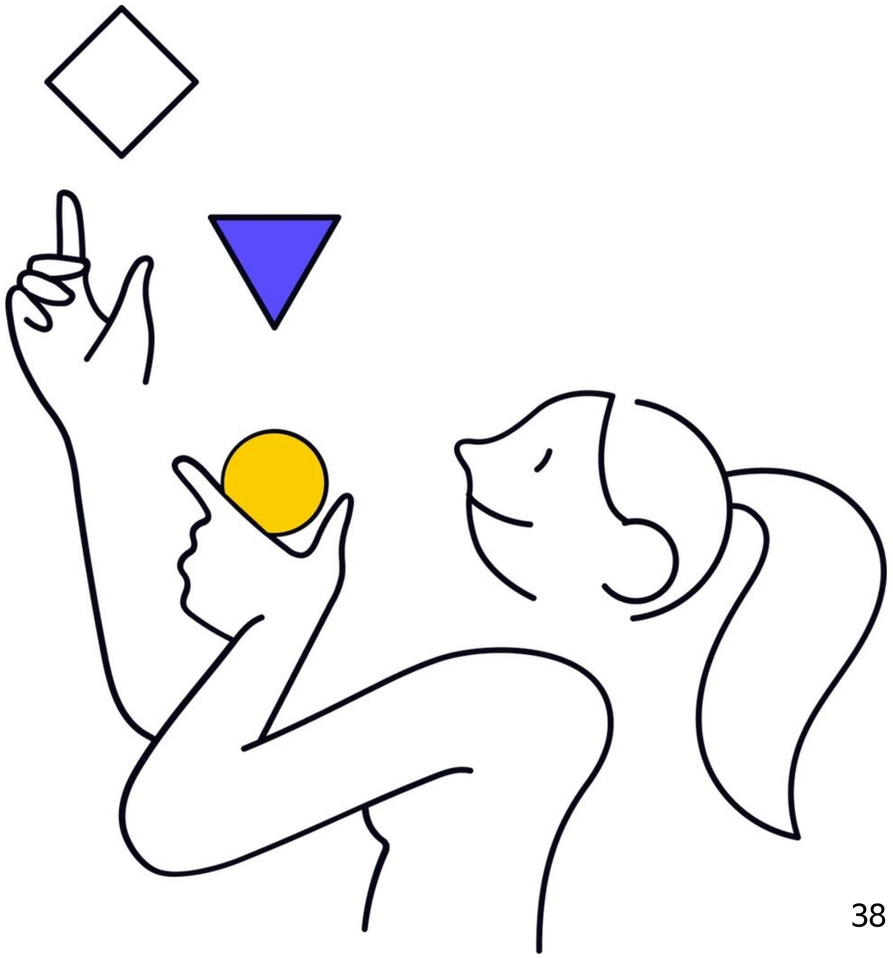
- Accuracy, Balanced accuracy
- Precision, Recall, F-score
- ROC curve, PR curve, AUC
- Confusion matrix

Next time

01 Support Vector Machines

02 Principal Component Analysis

03 Linear Discriminant Analysis



Q&A
Thanks for attention!

girafe ai

Radoslav Neychev