Introduction to Machine Learning
NPFL 054

http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká
hladka@ufal.mff.cuni.cz

Martin Holub
holub@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Outline

- Logistic regression
- Evaluation of binary classifiers

Binary classification
Decision boundary

A task of binary classification: $Y = \{0, 1\}$

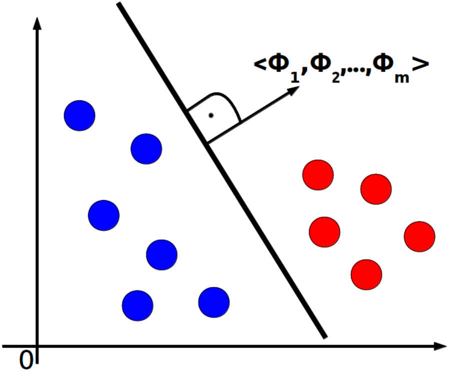
Decision boundary takes a form of function f and partitions a feature space into two sets, one for each class.

Binary classification
Hyperplane

Hyperplane is a linear decision boundary of the form

$$\Theta^\top \mathbf{x} = 0$$

where direction of $\langle \theta_1, \theta_2, \ldots, \theta_m \rangle$ is perpendicular to the hyperplane and θ_0 determines position of the hyperplane with respect to the origin



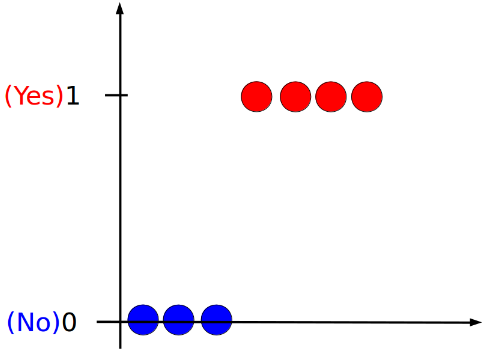
Binary classification
Hyperplane

- point if $m = 1$, line if $m = 2$, plane if $m = 3$, ...
- we can use hyperplane for classification so that

$$f(\mathbf{x}) = \begin{cases} 1 \text{ if } \theta_0 + \theta_1 x_1 + \cdots + \theta_m x_m \geq 0 \\ 0 \text{ if } \theta_0 + \theta_1 x_1 + \cdots + \theta_m x_m < 0 \end{cases}$$

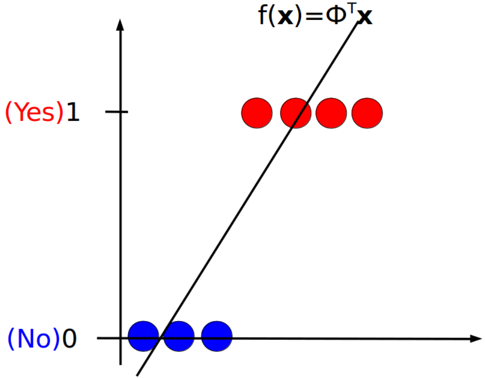
- **linear classifiers** classify examples using hyperplanes

Binary classification
Can we use linear regression?



Binary classification
Can we use linear regression?

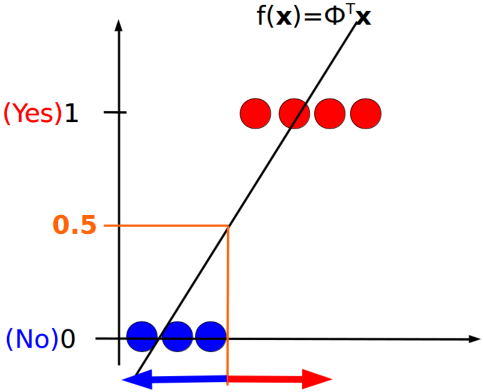
Fit the data with a linear function f



Binary classification
Can we use linear regression?

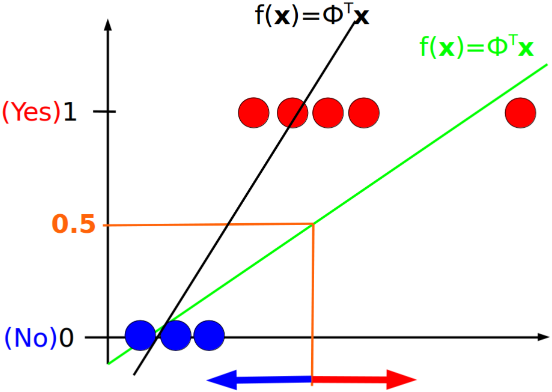
Classify

- if $f(\mathbf{x}) \geq 0.5$, predict 1
- if $f(\mathbf{x}) < 0.5$, predict 0



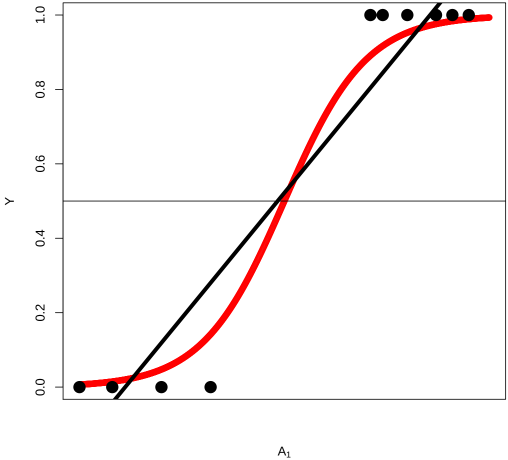
Binary classification
Can we use linear regression?

Add one more training instance

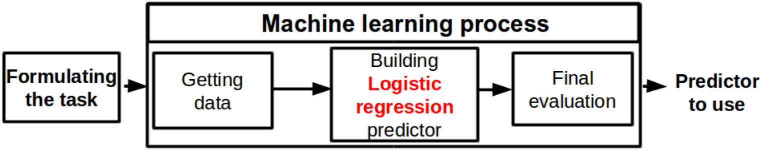


Binary classification
Can we use linear regression?

We are heading for the logistic regression algorithm.



Logistic regression

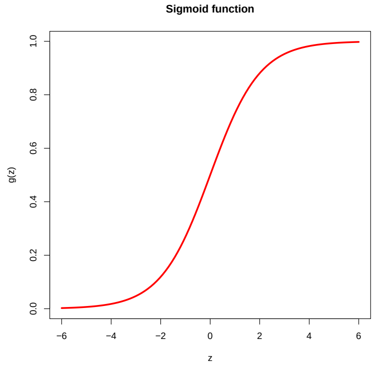


Logistic regression

Logistic regression is a classification algorithm.

Its target hypothesis f for a binary classification has a form of **sigmoid function**

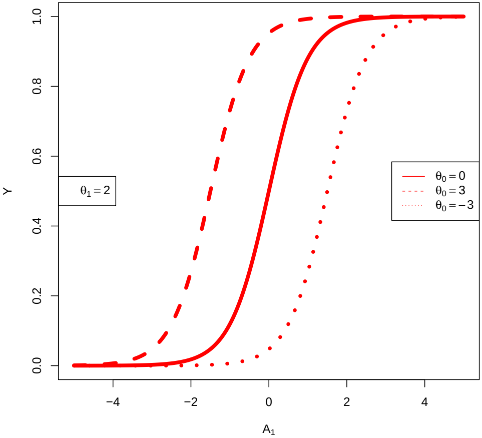
$$f(\mathbf{x};\Theta) = \frac{1}{1+e^{-\Theta^\top \mathbf{x}}} = \frac{e^{\Theta^\top \mathbf{x}}}{1+e^{\Theta^\top \mathbf{x}}}$$



- $g(z) = \frac{1}{1+e^{-z}}$
- $\lim_{z\to+\infty} g(z) = 1$
- $\lim_{z\to-\infty} g(z) = 0$

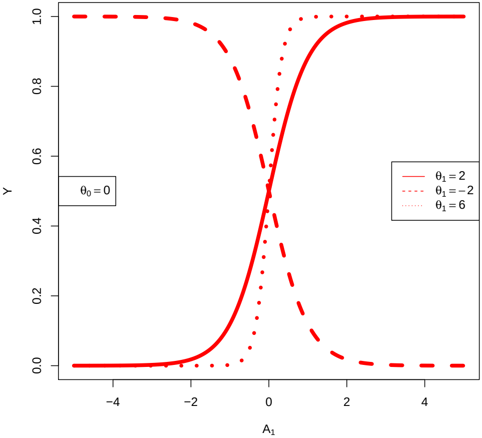
Logistic regression

$$f(\mathbf{x}; \Theta) = \frac{1}{1 + e^{-\theta_0 - \theta_1 x_1}}$$



Logistic regression

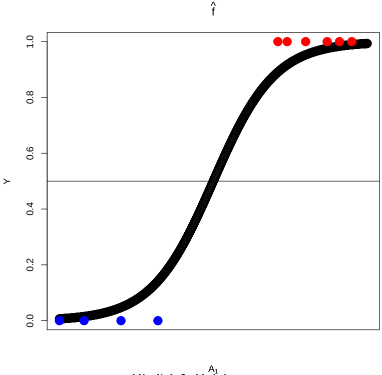
$$f(\mathbf{x}; \Theta) = \frac{1}{1 + e^{-\theta_0 - \theta_1 x_1}}$$



Logistic regression
Classification rule

Predict a target value using $\hat{f}(\mathbf{x}; \hat{\Theta})$ so that

- if $\hat{f}(\mathbf{x}; \hat{\Theta}) \geq 0.5$, i.e. $\hat{\Theta}^\top \mathbf{x} \geq 0$, predict 1
- if $\hat{f}(\mathbf{x}; \hat{\Theta}) < 0.5$, i.e. $\hat{\Theta}^\top \mathbf{x} < 0$, predict 0



Logistic regression
Derivation

Interpretation of $f(\mathbf{x};\Theta)$: it models the conditional probability $\Pr(y=1|\mathbf{x};\Theta)$

$$f(\mathbf{x};\Theta) = Pr(y=1|\mathbf{x};\Theta)$$

1. categorical attribute $Y = \{0,1\}$
2. $\cancel{y = \theta_0 + \theta_1 x_1 \cdots + \theta_m x_m}$, see above $\rightarrow$ model $\Pr(Y=y|\mathbf{x})$, e.g. $\Pr(Y=1|\mathbf{x})$
3. $\cancel{\Pr(Y=1|\mathbf{x}) = \theta_0 + \theta_1 x_1 \cdots + \theta_m x_m}$, see above
4. Model $\mathrm{odds}(\Pr(Y=1|\mathbf{x})) = \frac{\Pr(Y=1|\mathbf{x})}{\Pr(Y=0|\mathbf{x})} = \frac{\Pr(Y=1|\mathbf{x})}{1-\Pr(Y=1|\mathbf{x})} \in \langle 0, +\infty)$

Odds, odds ratio

$\text{odds} = \Pr(\text{success}) / \Pr(\text{failure})$

Titanic example

```
> d <- read.csv("train.csv")
> attach(d)
> table(Sex, Survived)
        Survived
Sex        0   1
  female  81 233
  male   468 109
> detach()
```

- the odds of surviving for male:
  $\Pr(\text{Survived} = 1 | \text{Sex} = \text{male}) / \Pr(\text{Survived} = 0 | \text{Sex} = \text{male}) = \frac{109}{486} = 0.23$
- the odds of surving for female:
  $\Pr(\text{Survived} = 1 | \text{Sex} = \text{female}) / \Pr(\text{Survived} = 0 | \text{Sex} = \text{female}) = \frac{233}{81} = 2.88$
- the ratio of the odds for female to the odds for male $2.88/0.23 = 12.52$

Logit

5. Transform $\langle 0, +\infty)$ to $(-\infty, +\infty)$: model

$$\mathrm{logit}(\Pr(Y = 1|\mathbf{x})) = \ln(\mathrm{odds}(\Pr(Y = 1|\mathbf{x}))) = \ln(\frac{\Pr(Y = 1|\mathbf{x})}{1 - \Pr(Y = 1|\mathbf{x})})$$

6. Use linear regression

$$\ln(\frac{\Pr(Y = 1|\mathbf{x})}{1 - \Pr(Y = 1|\mathbf{x})}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_m x_m$$

i.e.,

$$\Pr(Y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\theta_0 - \theta_1 x_1 - \cdots - \theta_m x_m}}$$

$$\boxed{f(\mathbf{x_i}; \Theta) = \Pr(Y_i = 1|\mathbf{x_i}; \Theta) = \frac{1}{1+e^{-\Theta^\top \mathbf{x_i}}}}$$

Parameter interpretation
Binary features

- Use $\mathrm{female} = \{1, 0\}$ instead of $\mathrm{Sex} = \{\mathrm{female}, \mathrm{male}\}$
- in Linear regression $y = \theta_0 + \theta_1 * \mathrm{female}$
  - θ_0 is the average y for male
  - $\theta_0 + \theta_1$ is the average y for female
  - θ_1 is the average difference in y between female and male
- in Logistic regresion $p = \Pr(Survive = 1|\mathbf{x}, \Theta)$, $\ln \frac{p}{1-p} = \theta_0 + \theta_1 * \mathrm{female}$
  - **If** $\mathrm{female} == 0$
    - $p = p_1 \rightarrow \ln(\frac{p_1}{1-p_1}) = \theta_0 \rightarrow \frac{p_1}{1-p_1} = e^{\theta_0}$
    - the intercept θ_0 is the log odds for men
  - **If** $\mathrm{female} == 1$
    - $p = p_2 \rightarrow \frac{p_2}{1-p_2} = e^{\theta_0+\theta_1}$
    - odds ratio $= \frac{p_2}{1-p_2} / \frac{p_1}{1-p_1} = e^{\theta_1}$
    - the parameter θ_1 is the log odds ratio between female and male

Parameter interpretation
Numerical features

- θ_i gives an average change in $\mathrm{logit}(f(\mathbf{x}))$ with one-unit change in A_i holding all other features fixed

Parameter estimates

- **Loss function**

$$\mathrm{L}(\Theta) = -\sum_{i=1}^{n} y_i \log \mathrm{P}(y_i|\mathbf{x_i};\Theta) + (1-y_i)\log(1-\mathrm{P}(y_i|\mathbf{x_i};\Theta))$$

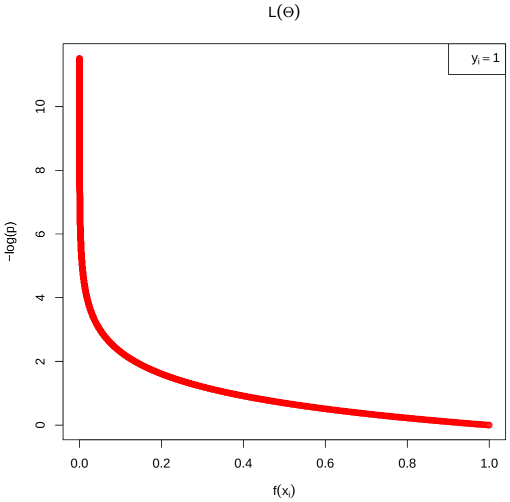
See Maximum Likelihood Principle for derivation of this loss function.

- **Optimization problem**

$$\Theta^\star = \mathrm{argmin}_\Theta \mathrm{L}(\Theta)$$

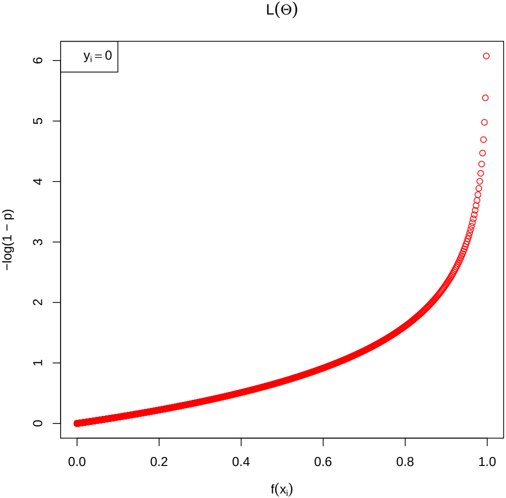
Parameter estimates

$$\mathrm{L}(\Theta) = -\sum_{i=1}^{n} y_i \log \mathsf{P}(y_i|\mathbf{x_i};\Theta) + (1-y_i)\log(1-\mathsf{P}(y_i|\mathbf{x_i};\Theta))$$



Parameter estimates

$$\mathrm{L}(\Theta) = -\sum_{i=1}^{n} y_i \log \mathrm{P}(y_i|\mathbf{x_i};\Theta) + (1-y_i)\log(1-\mathrm{P}(y_i|\mathbf{x_i};\Theta))$$



Parameter estimates
Gradient Descent Algorithm

repeat until convergence $\{$

$$\Theta^{K+1} := \Theta^{K} - \alpha \nabla f(\Theta^{K})$$

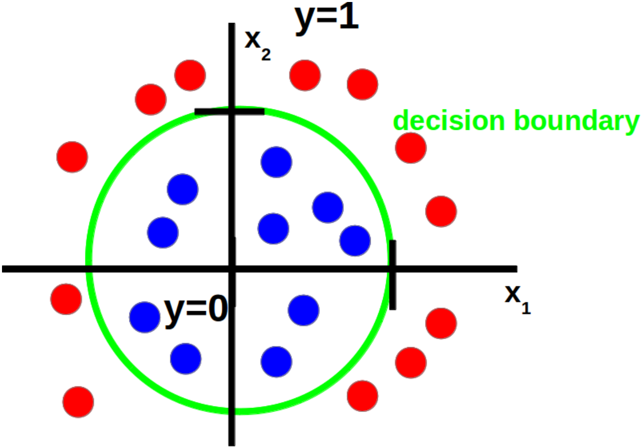
$\}$

– α is a positive step-size hyperparameter

I.e. simultaneously update θ_j, $j = 1, \ldots, m$

$$\theta_j^{K+1} := \theta_j^{K} - \alpha \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i; \Theta^{K}) - y_i) x_{ij}$$

Non-linear decision boundary



Non-linear decision boundary

- For hyperplane: $f(\mathbf{x}) = g(\theta_0 + \theta_1 x_1 + \cdots + \theta_m x_m)$
- Let $f(\mathbf{x}) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$ (a higher degree polynomial)
- Assume $\theta_0 = -1$, $\theta_1 = 0, \theta_2 = 0$, $\theta_3 = 1, \theta_4 = 1$
- Predict $y = 1$ if $-1 + x_1^2 + x_2^2 \geq 0$, i.e. $x_1^2 + x_2^2 \geq 1$

Logistic regression
Summary

Classification of $\mathbf{x}$ by $\hat{f}^\star$

1. Project $\mathbf{x}$ onto $\hat{\Theta}^\star$ to convert it into a real number z in the range $(-\infty, +\infty)$
   - i.e. $z = {\hat{\Theta}^\star}^\top \mathbf{x}$
2. Map z to the range $\langle 0, 1 \rangle$ using the sigmoid function $g(z) = 1/(1 + e^{-z})$
3. Classify $\mathbf{x}$ using a classification rule

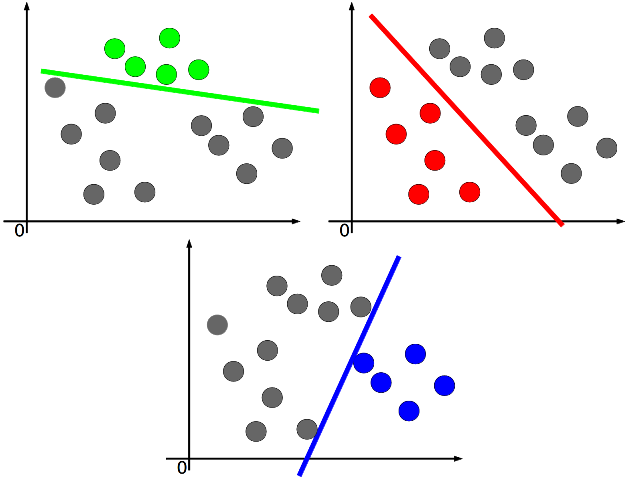
Multi-class classification

$|Y| = N$, $N \geq 3$

- **One-to-all**
  – train N binary classifiers f_k for the pair k-th class and $\{1, \cdots, N\} \setminus \{k\}$ classes
  – classify $\mathbf{x}$ into the class $k^\star = \mathrm{argmax}_k f_k(\mathbf{x})$
- **One-to-one**
  – train $\binom{N}{2}$ binary classifiers f_i for each pair of classes
  – classify $\mathbf{x}$ into the class $k^\star = \max_{k=1,\dots,N} \sum_{i=1}^{\binom{N}{2}} \delta(f_i(\mathbf{x}) = k)$

Logistic regression
Multi-class classification

One-to-all



Evaluation of binary classifiers
Confusion matrix

Confusion matrix is a square matrix indexed by all possible target class values.

Task: Assign the correct sense of the word *line* in a sentence.

```
** Comparing the predicted values with the true senses **

           Prediction
Truth       cord division formation phone product text
  cord       268        3        10     7       9    6
  division     3      280         1     2       5    3
  formation   13        3       225     4      19    4
  phone       25        5         2   293      12   10
  product     51       10        39    32    1442   72
  text        12        1         7     4      28  262
```

Correctly predicted examples are displayed on the diagonal.

Evaluation of binary classifiers
Confusion matrix

In binary classification tasks examples are sometimes regarded as divided into two disjoint subsets:

- **positive examples** – "to be retrieved" (ones)
- **negative examples** – "not to be retrieved" (zeros)

```
### Example confusion matrix for binary classification
> table(test.true, test.pred)
        prediction
           0   1
 true  0 580  69
       1  37 144
>
```

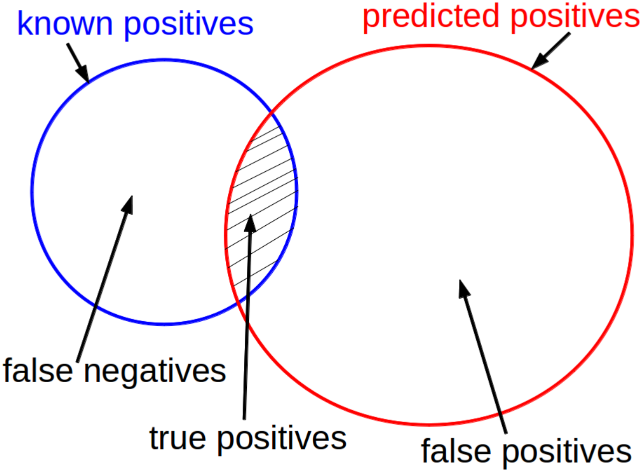
Evaluation of binary classifiers
Confusion matrix

		Predicted class	
		Positive	**Negative**
True class	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Explanation

- **'Trues'** are examples correctly classified
- **'Falses'** are examples incorrectly classified
- **'Positives'** were predicted as positives (correctly or incorrectly)
- **'Negatives'** were predicted as negatives (correctly or incorrectly)

Proportion of correctly predicted test examples



Basic performance measures

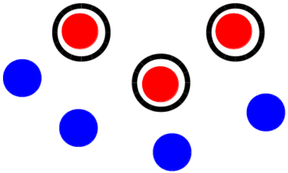
Measure	Formula
Precision	TP/(TP+FP)
Recall/Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)
1-Specificity (FPR)	FP/(TN+FP)
Accuracy	(TP+TN)/(TP+FP+TN+FN)

Very often you need to **combine both good precision and good recall**. Then you usually use **balanced F-score**, so called **F-measure**

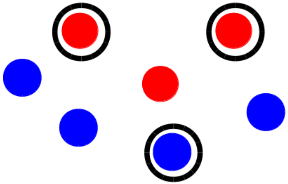
$$\boxed{F = 2\,\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}}$$

Sensitivity vs. Specificity

Perfect classifier – no error

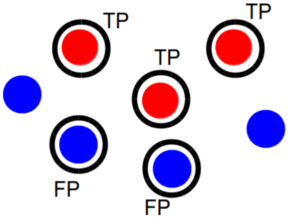


Reality – errors

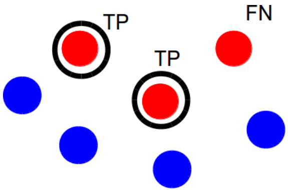


Sensitivity vs. Specificity

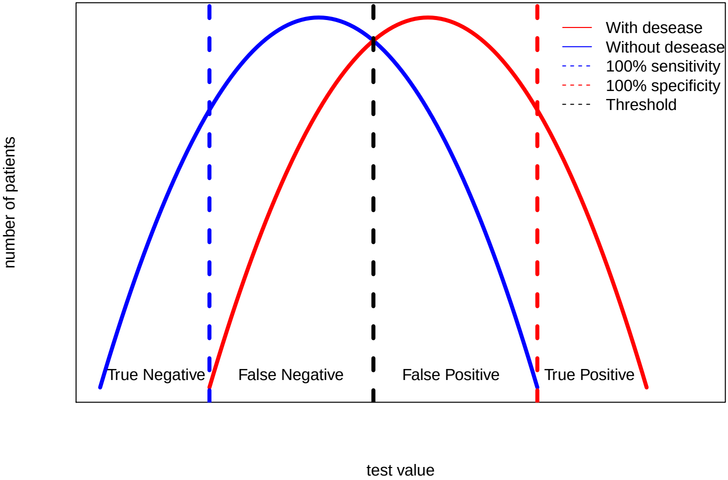
100% sensitive classifier



100% specific classifier

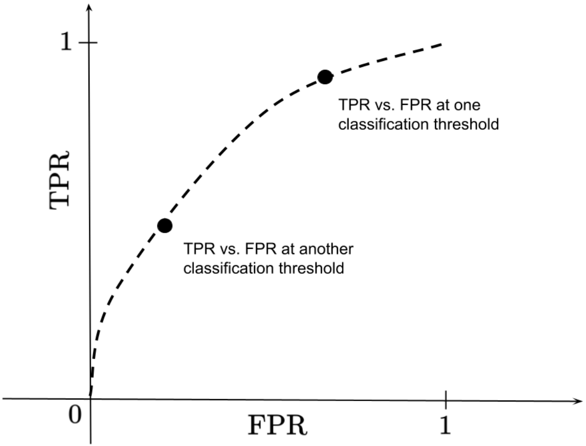


Sensitivity vs. specificity



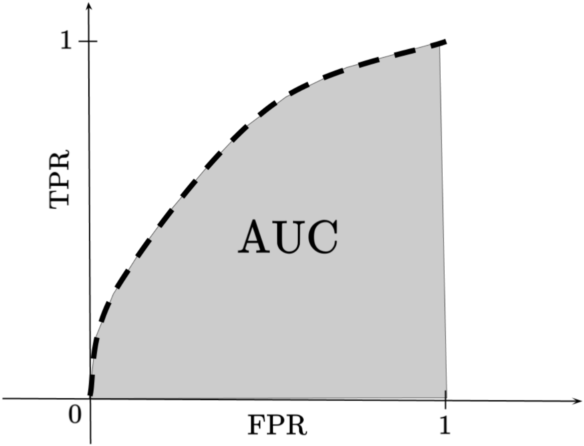
ROC curve

An **ROC curve** plots True Positive Rate vs. False Positive Rate at different classification thresholds (see p. 6).

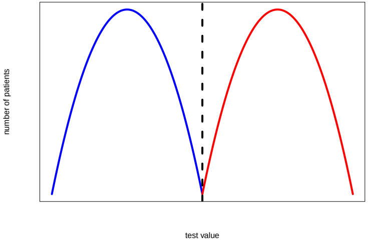


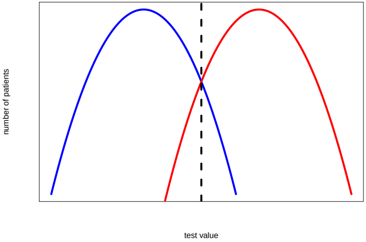
Evaluation of binary classifiers
AUC measure

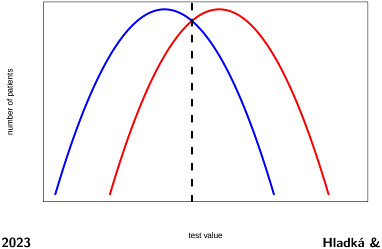
Area Under ROC (= AUC)
is a measure of how good is a distinguishing property of classifier

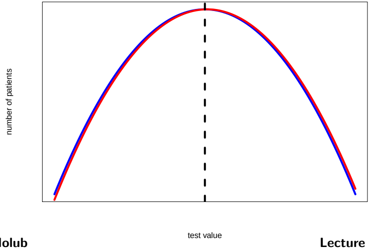


Evaluation of binary classifiers
ROC & AUC

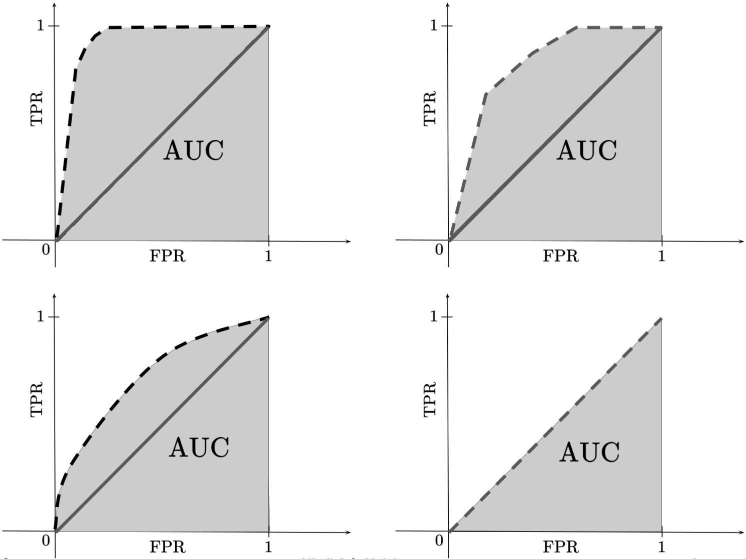








Evaluation of binary classifiers
ROC & AUC



Summary of Examination Requirements

- Decision boundary, hyperplane, classification rule
- Logistic regression, sigmoid function, probabilistic formulation
- Multi-class classification
- Confusion matrix for binary classification
- Basic performance measures, ROC curve, AUC

For more details refer to

- Ng Andrew, Machine Learning online at Stanford (https://class.coursera.org/ml-2012-002/class/index)
- Hastie, T. et al. *The Elements of Statistical Learning*. Springer, 2009, (http://www-stat.stanford.edu/~tibs/ElemStatLearn/), Section 4.4