Universiteti i Prishtinës "Hasan Prishtina" Kosovë

# Hyrje në shkencën e të dhënave

Pjesa 9 – Mësimi i mbikqyrur (Supervised Learning)

Prof. Asoc. Dr. Ermir Rogova

Universiteti i Prishtinës "Hasan Prishtina" Kosovë

# What is supervised learning?

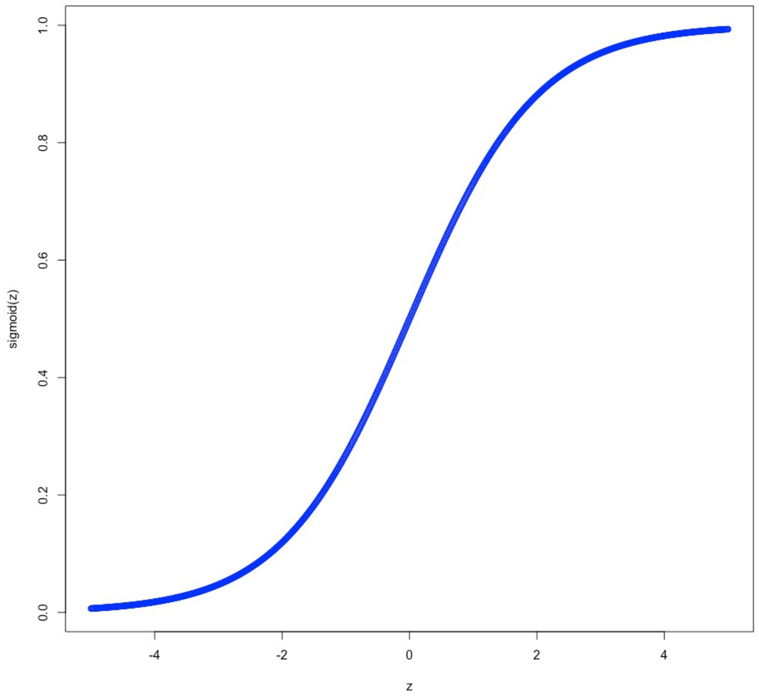
- Learning from data when we have correct labels / outcome values
- Example: knowing the relationship between size of tumor and cancer (yes/no)
- A major area: classification

# What we will cover

- Logistic regression
- Softmax regression
- kNN
- Decision tree
- Random forest
- Naïve Bayes
- Support Vector Machine (SVM)

# Logistic regression

- Use of sigmoid function to turn a continuous predicted value to something bound between 0 and 1, and then to Class-1 (below 0.5) and Class-2 (above 0.5).

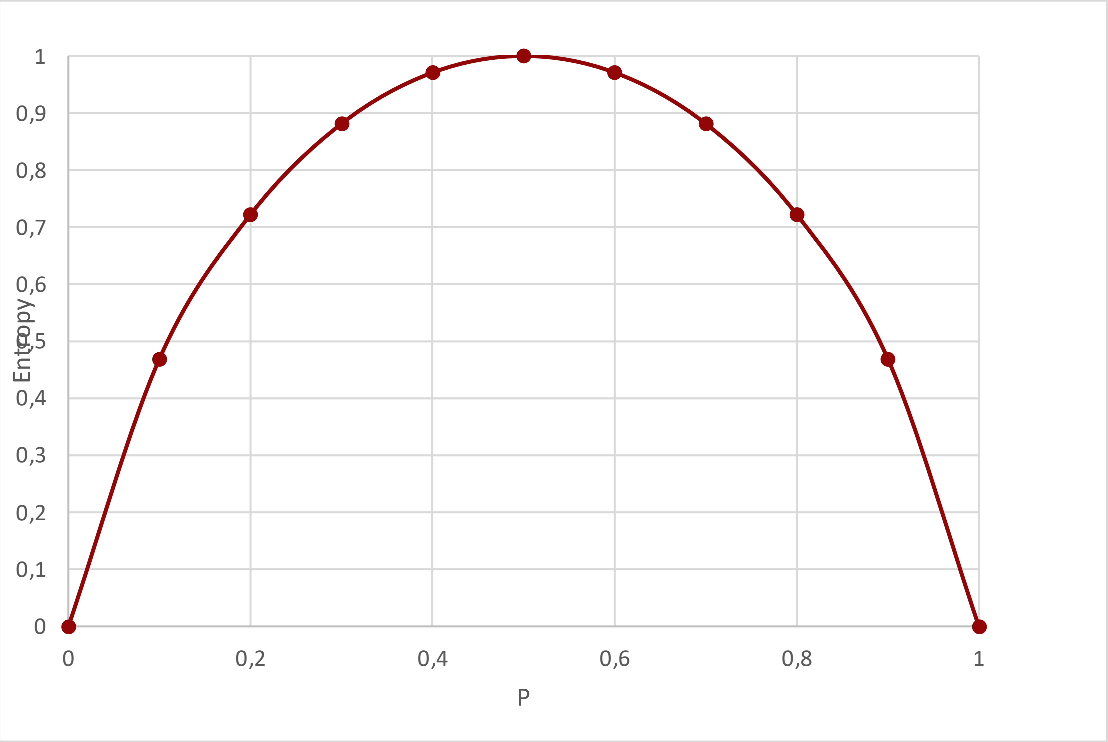


# kNN (k nearest neighbor)

- As in the general problem of classification, we have a set of data points for which we know the correct class labels.
- When we get a new data point, we compare it to each of our existing data points and find similarity.
- Take the most similar k data points (k nearest neighbors).
- From these k data points, take the majority vote of their labels. The winning label is the label/class of the new datapoint.

Universiteti i Prishtinës "Hasan Prishtina" Kosovë

# Entropy and information gain



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$IG(A, B) = Entropy(A) - Entropy(A, B)$$

# Decision tree

- Step 1: Calculate Entropy of the target or class variable
- Step 2: The dataset is then split on the different attributes into smaller sub-tables. The entropy for each sub-table is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain or decrease in entropy.
- Step 3: Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

# Random forest

- Problem with decision tree: could overfit the data, making it difficult to do well on new data
- A solution: grow many decision trees that are randomly paralyzed, and have them vote for an outcome = random forest

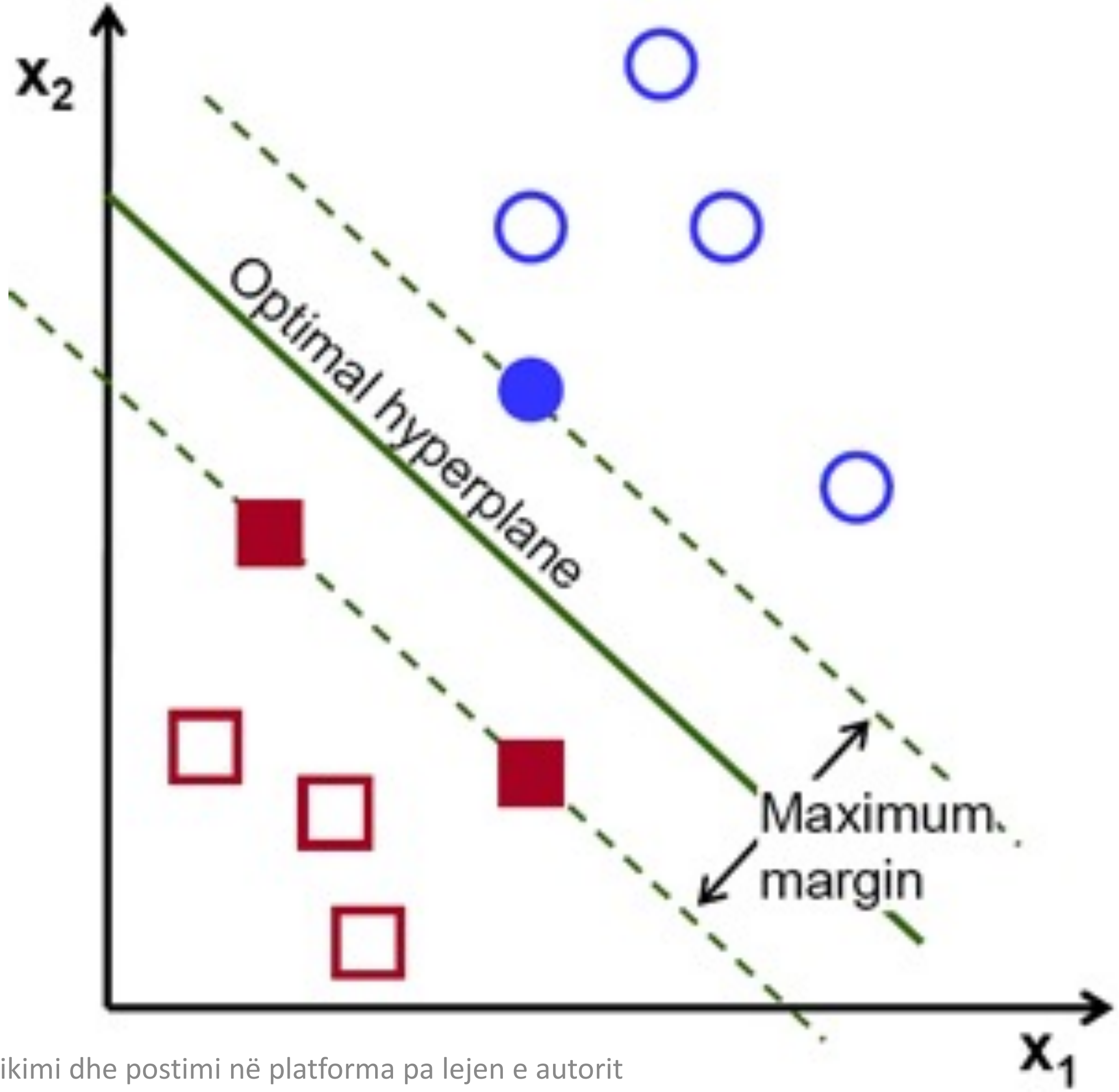
# Naïve Bayes

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood, which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

# Support Vector Machine (SVM)

- Uses nonlinear mapping to transform the original training data into a higher dimension.
- Within this new dimension, it searches for the linear optimal separating hyperplane.



Universiteti i Prishtinës "Hasan Prishtina" Kosovë

# Summary

- Supervised learning algorithms use a set of examples from previous records that are labeled to make predictions about future.
- Two main branches: regression and classification
- Classification techniques:
  - Logistic regression (two-class problems)
  - Softmax regression (multi-class problems)
  - kNN
  - Decision tree
  - Random forest
  - Naïve Bayes
  - SVM

Universiteti i Prishtinës "Hasan Prishtina" Kosovë

# Temat e javës së dhjetë

- Unsupervised Learning
- Agglomerative Clustering
- Divisive Clustering
- Expectation Maximization (EM)
- Introduction to Reinforcement Learning

# Pyetje???