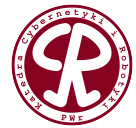
Selected topics in Artificial Intelligence
Machine learning

Wojciech Domski

Chair of Cybernetics and Robotics,
Wrocław University of Science and Technology

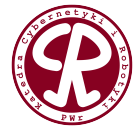
Presentation compiled for taking notes during lecture

Wrocław University
of Science and Technology



Outline

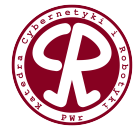
Wrocław University
of Science and Technology



Classifier

Classifier is a tool which allows to classify a sample to a class given learning set with the feature and label association.

Wrocław University of Science and Technology



Bayes Theorem (1/2)

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{1}$$

(1) gives probability of some event A (hypothesis) happening given that a different event B (evidence) has occurred.
$P(A|B)$ is so-called a posterior probability.
$P(B|A)$ is conditional probability of event B given event A.
$P(A)$ probability of event A, a prior probability.
$P(B)$ probability of event B happening independent from event A.
Thus (1) can be expressed as

$$posterior = \frac{prior \cdot likelihood}{evidence} \tag{2}$$

Wrocław University of Science and Technology

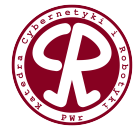
Bayes Theorem (2/2)

Let's assume that the B (evidence or features) are independent then we can assume that

$$P(A|B) = \frac{1}{k} P(A) \prod_{i=1}^{n} P(B_i|A) \tag{3}$$

where k is a scaling factor which is dependant only on B and is constant when values of features are known.

Wrocław University of Science and Technology



Types of Bayes Classifier (1/2)

Following types of Bayes Classifier can be enumerated.

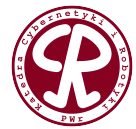
Multimodal

This type of classifier is used for discrete type of data. In the problem of classification of multiple classes this type of Bayes Classifier is being used.

Gaussian

Assumes that the data has normal type of distribution. Probability density function is so-called bell curve with mean value of μ and standard deviation of σ.

Wrocław University of Science and Technology

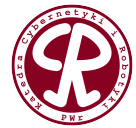


Types of Bayes Classifier (2/2)

Bernoulli

Bernoulli distribution is a discrete probability distribution of some random variables that can take either 1 (with probability p) or 0 (with probability $q = 1 - p$), thus it is used to classify samples that are binary.

Wrocław University of Science and Technology



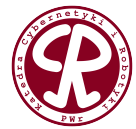
Learning process (1/3)

The training process of Naive Bayes Classifier can be broke down into following steps:

1. Divide – separate data set into subsets by given classes.
2. Dataset statistic – calculate mean and standard deviation of the whole set.
3. Class statistic – calculate statistic as in previous step but per class.
4. Calculate probability for each class.

To calculate $P(A|B)$ (3) can be used since we assume that features B are independent.

Wrocław University of Science and Technology



Learning process (2/3)

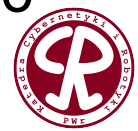
Let's assume that we have two classes ($n = 2$) and a sample is classified based on two features ($k = 2$).

1. Separate dataset according to classes.
2. For each separate class calculate mean and standard deviation for each feature. Create matrix of mean values $M_{n\times k}$ and matrix of standard deviations $S_{n\times k}$.

Estimation can be decomposed into following steps

1. For each class calculate probability of a sample belonging to the class $P(A|B)$.
2. $P(A|B) = \frac{1}{k}\mathbf{P}(\mathbf{A})\prod_{i=1}^{n} P(B_i|A)$
   Calculate probability of the sample belonging to given class by initiating it with the share of samples belonging to the given class which approximates $P(A)$.

Wrocław University of Science and Technology



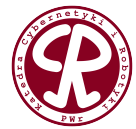
Learning process (3/3)

3. $P(A|B) = \frac{1}{k} P(A) \prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{P(B_i|A)}$ Multiply above probability with the estimation of the probability per feature for given sample.

The highest score represents the estimated class. To calculate probability of belonging to a given class softmax formula can be used which gives overall share for each class, i.e.

$$s_i = \frac{x_i}{\sum_{i=1}^{n} x_i}. \tag{4}$$

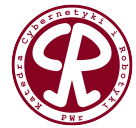
Wrocław University of Science and Technology



Decide and split (1/4)

Decision tree is a technique which allows to classify as well as perform regression given a set of features. With every step a split is performed in reference to one of the features based on calculated coefficient.
To select a feature which provides the most effective split of the data a metrics needs to be introduced. There are a selection of methods which can be used for this purpose.

Wrocław University
of Science and Technology



Decide and split (2/4)

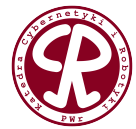
Entropy

Entropy is a measure to quantify randomness in the data.

$$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{5}$$

Let's assume that there is a set of data with two types of samples, e.g. cats and dogs. Share of cats is 20%. Then the entropy is equal to 0.72.

Wrocław University of Science and Technology



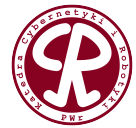
Decide and split (3/4)

Gini impurity index

It is a metric to measure impurity of the data by checking misclassification of a random sample classified according to the distribution of labels.

$$G(p) = 1 - \sum_{i=1}^{n} p_i^2. \tag{6}$$

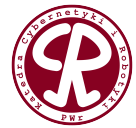
Wrocław University of Science and Technology



Decide and split (4/4)

The are other methods like Information gain, Variance reduction or Measure of "goodness" which all base on different concepts but allow to achieve the same result, i.e. deciding which feature can be used for a split.

Wrocław University of Science and Technology

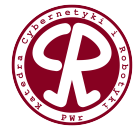


Tree building (1/1)

Besides metrics which allows to make a decision where to perform the split a learing algorithm is needed to build a decision tree. The decision tree is a graph where in each node a feature from set of features is assessed. When a leaf of a tree is reached a class is assigned for the sample.
Following tree building algorithms are commonly used:

- ID 3 (Iterative dichotomiser),
- C4.5,
- CART (Classification and Regression Tree).

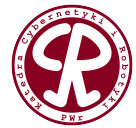
Wrocław University
of Science and Technology



Random forest/Random trees

Decision tree is a technique allowing to perform classification or regression. However, the effectiveness of a single decision tree can be increased.
Random forest is a set of decision trees overcoming weakness of decision tree – over fitting.
For classification problem the most common predicted class is chosen as the prediction, thus majority voting rule is being used. For regression the prediction is calculated as an average of all predicted values coming from separate decision trees.

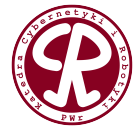
Wrocław University
of Science and Technology



Tree pruning

As mentioned before a single decision tree is prone to over fitting problem. Instead of using multiple decision trees a different techniques can be used – tree pruning.
The goal of tree pruning is to remove part of the decision tree which do not contribute to solution. This action allows to decrease the complexity of the graph as well as reducing the over fitting problem.

Wrocław University
of Science and Technology



Support Vector Machine

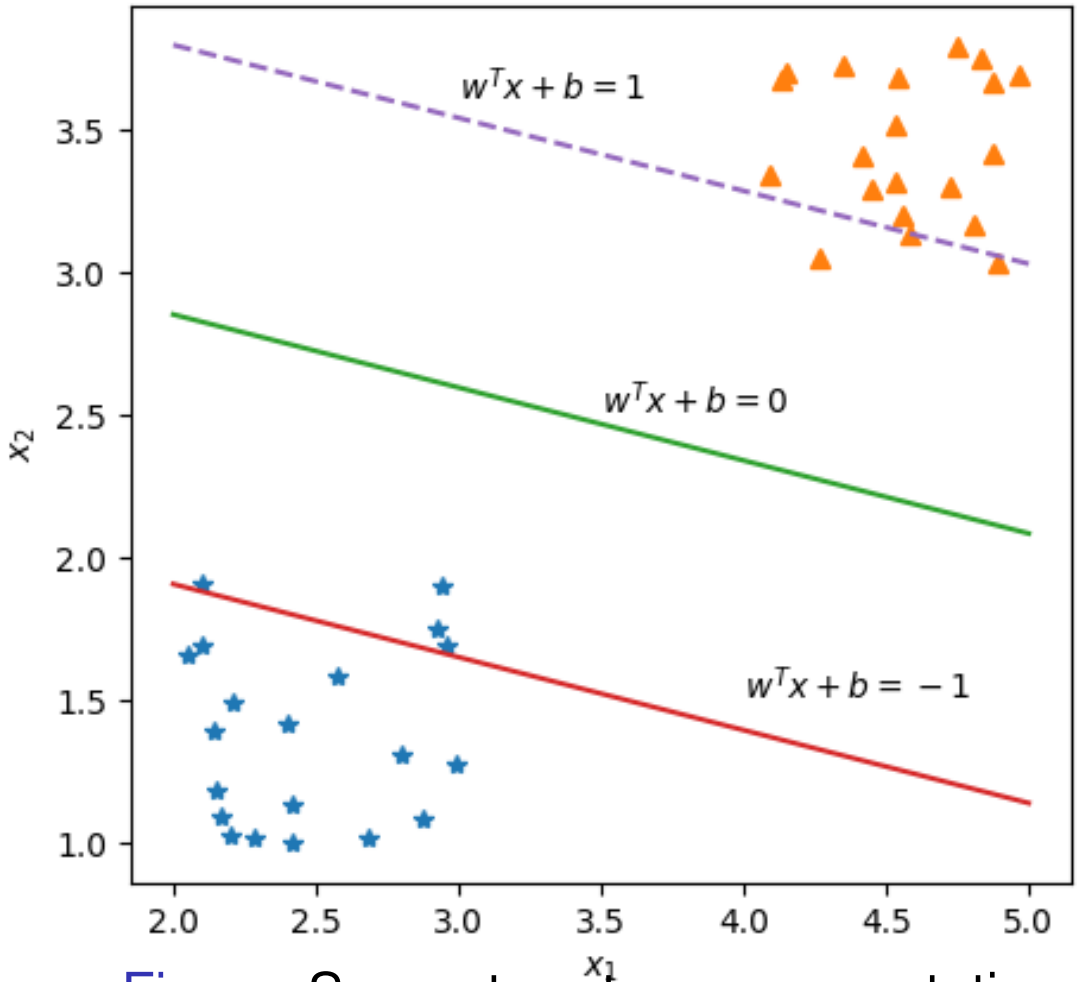


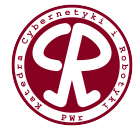
Figure: Support vectors representation

Support Vector Machine (1/5)

Support Vector Machine, similarly to Decision tree, can perform classification or regression. However, SVM supports binary classification which means that it can distinguish between two classes at a time. In order to handle multi class (multi label) data two methods can be used:

- *One to One*
  Separate classifier is created between all pairs of classes. This lead to $\frac{m(m-1)}{2}$ individual classifiers. During this method only samples for given two classes are considered, thus during training other samples are not included.

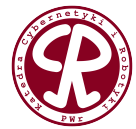
Wrocław University of Science and Technology



Support Vector Machine (2/5)

- *One to Rest*
  This approach allows to create n separate classifiers where n is considered to be number of classes themselves. During training samples are divided into two sets. One set includes representatives from a given i^{th} class while second set includes all other samples.

Wrocław University
of Science and Technology



Support Vector Machine (3/5)

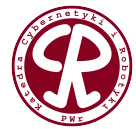
SVM minimizes cost function given as

$$J(w) = \frac{1}{2}||w||^2 + C\left(\frac{1}{N}\sum_{i=1}^{N}\max(0, d_i)\right) \tag{7}$$

where C regulates width of the margin (distance between support vectors). With larger C the margin shrinks. d_i is distance between sample x_i and the border line given as

$$d_i = 1 - y_i(wx_i + b).$$

Wrocław University of Science and Technology



Support Vector Machine (4/5)

To simplify computations the bias b can be incorporated inside weights vector w but at the same time the feature vector x has to be extended so

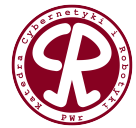
$$wx_i + b = \bar{w}\bar{x}_i \tag{8}$$

where $\bar{w} = (b, w)^T$ and $\bar{x}_i = (1, x_i)^T$.
In order to train the classifier (find the best set of weights w) gradient of $J(w)$ has to be calculated.

$$\nabla J(w) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} w & \text{if } max(0, d_i) = 0 \\ w - Cy_i x_i & \text{otherwise} \end{cases}. \tag{9}$$

Through numerical optimization, e.g. using SGD (stochastic gradient descent) estimation of optimal weights values $\tilde{w}$ can be calculated.

Wrocław University of Science and Technology



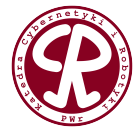
Support Vector Machine (5/5)

To determine the classification result sign function can be used to evaluate formula

$$y^* = sign(wx + b). \tag{10}$$

Through this expression binary classification can be performed.

Wrocław University of Science and Technology



Kernel (1/3)

SVM uses following function

$$y(x) = w^T x + b \tag{11}$$

to predict the output. However, it can be rewritten as

$$y(x) = w^T x + b = \sum_{i=1}^{n} \left(a_i x^T x \right) + b. \tag{12}$$

The dot product $x^T x$ can be replaced and generalised as a kernel function

$$k(x, x) = \Phi(x) \cdot \Phi(x). \tag{13}$$

A kernel is generally an inner product, thus a dot product in particular.

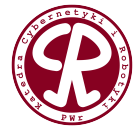
Wrocław University of Science and Technology

Kernel (2/3)

In turn, this allows to replace feature vector x with the output of $\Phi(x)$ while the dot product is replaced with the kernel function $k(x, x)$.
Using kernel function it allows us to make predictions according to the rule

$$y(x) = w^T x + b = \sum_{i=1}^{n} (a_i k(x, x)) + b. \tag{14}$$

Wrocław University of Science and Technology

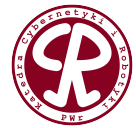


Kernel (3/3)

A kernel function is valid if and only if the Mercel's condition is met. Let's express kernel function as a pairwise dot product yielding a square matrix $K^{n \times n}$. The K matrix should be symmetric and positive semi-defined, thus

$$\begin{cases} K = K^T \\ det(K) \geq 0 \end{cases} . \tag{15}$$

Wrocław University of Science and Technology



Kernel types (1/3)

Different types of kernels can be distinguished.

Polynomial kernel

Two kinds of polynomial kernels can be distinguished – homogeneous and inhomogeneous. Homogeneous kernel can be expressed as

$$K(x, y) = (x^T y)^p. \tag{16}$$

For $p = 1$ it degenerates to a linear expression.
Inhomogeneous kernel has following formula

$$K(x, y) = (x^T y + c)^p \tag{17}$$

where $c \geq 0$ is a constant. As it can be seen for $c = 0$ inhomogeneous kernel degenerates to homogeneous kernel. Usually, p should be equal to the number of features.

Kernel types (2/3)

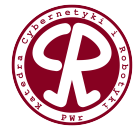
Gaussian kernel

$$K(x, y) = e^{-\frac{||x-y||^2}{2\sigma^2}} \tag{18}$$

With high σ values can cause over fitting problem.
Sometimes Gaussian kernel can be referred to as Radial Basis Function (RBF) and is usually expressed as

$$K(x, y) = e^{-\gamma||x-y||^2} \tag{19}$$

Wrocław University of Science and Technology



Kernel types (3/3)

Sigmoid kernel

$$K(x, y) = \tanh(\gamma(x^T y) + c) \tag{20}$$

Wrocław University
of Science and Technology

Dual problem (1/2)

For linear kernel function solving SVM is considered to be a primal problem. We need to optimize $n+1$ variables where n is number of features while 1 accounts for a bias. To solve linear SVM, thus the primal problem optimization technique such as (SGD) stochastic gradient descent can be used.

$$minimize \frac{1}{n}\sum_{i=1}^{n} d_i + \lambda||w||^2 \tag{21}$$

subject to

$$y_i(w^T x_i - b) \geq 1 - d_i \tag{22}$$

where $d_i = max(0, 1 - y_i(w^T x_i - b))$ and $d_i \geq 0 \forall i$.
The primal problem performs minimization of a given cost function subject to certain constrains.

Dual problem (2/2)

$$maximize\ f(c) = \sum_{i=1}^{n} c_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i c_i (x_i^T x_j) y_j c_j \tag{23}$$

subject to

$$\begin{cases} \sum_{i=1}^{n} c_i y_i = 0 \\ 0 \leq c_i \leq \frac{1}{2n\lambda} \forall i \end{cases} \tag{24}$$

To solve above quadratic programming can be used so c_i is in following relation with weights

$$w = \sum_{i=1}^{n} c_i y_i x_i. \tag{25}$$

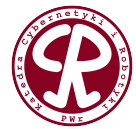
Dual problem tries to perform maximization while being subject to other constrains. Additionally, each feasible solution of a primal problem is an upper bound solution for the dual problem.

Wrocław University of Science and Technology

Non linear kernel for primal problem (1/7)

Can we use non linear kernel for primal problem ?

Wrocław University
of Science and Technology



Non linear kernel for primal problem (2/7)

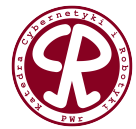
Let's consider following feature vector

$$x = (x_1, x_2). \tag{26}$$

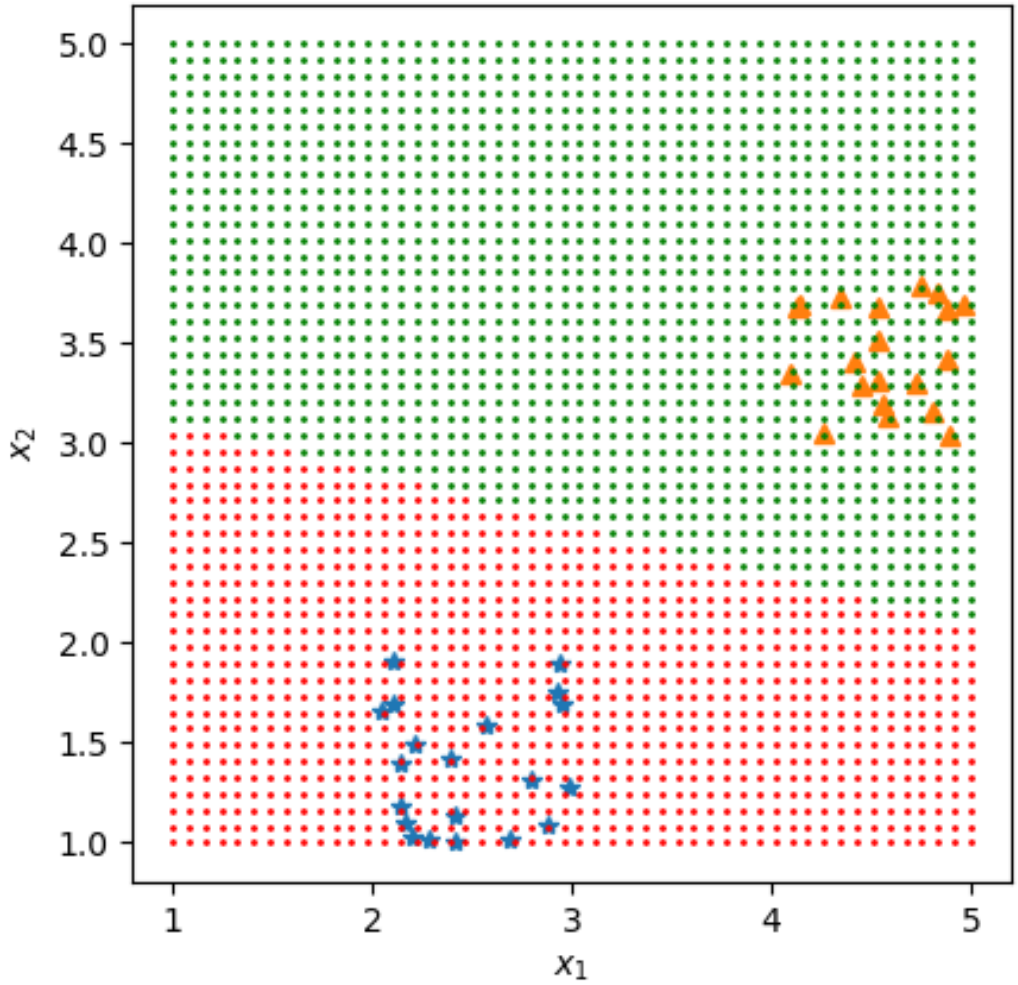
To solve and find weights w we can use SGD. Additionally to above we can extend the feature vector to account for bias, thus

$$x = (1, x_1, x_2) \tag{27}$$

Wrocław University
of Science and Technology



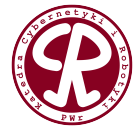
Non linear kernel for primal problem (3/7)



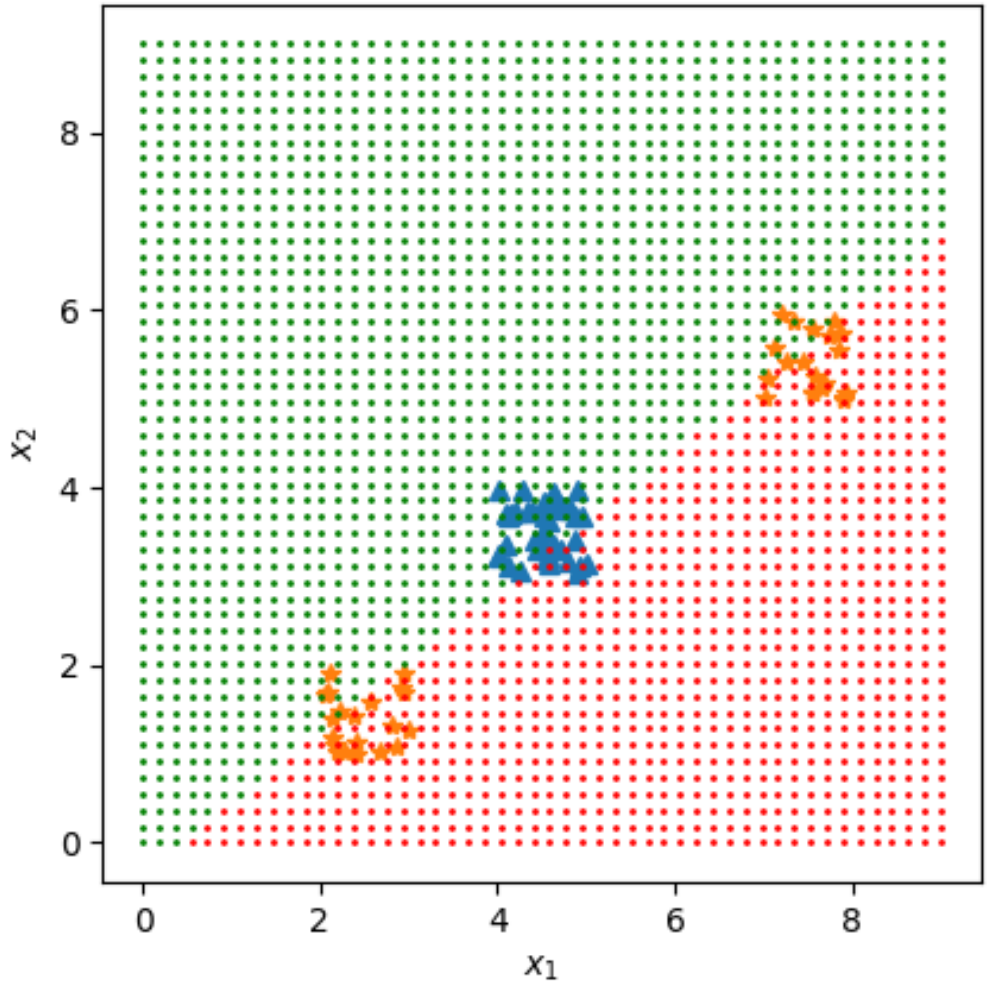
Non linear kernel for primal problem (4/7)

Let's consider a set of features which can not be separated with linear kernel

Wrocław University of Science and Technology



Non linear kernel for primal problem (5/7)



Non linear kernel for primal problem (6/7)

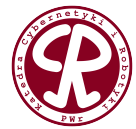
To achieve classes separation the feature space has to be extended, thus previous feature vector

$$x = (1, x_1, x_2) \tag{28}$$

can be extended to polynomial representation

$$x = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2) \tag{29}$$

Wrocław University of Science and Technology



Non linear kernel for primal problem (7/7)