Classic methods

November 18, 2016

Agenda

1. Metrics
2. Support vector machines
3. Decision trees, random forest
4. Clusterisation. K-means
5. Dimension reduction. PCA

Useful resources

1. Coursera. Machine learning (Andrew Ng)
2. Coursera. Введение в машинное обучение (Higher School of Economics)

Metrics

1. Accuracy
2. Precision, recall
3. Confusion matrix
4. F-measure
5. PR-curve
6. Area under PR-curve
7. Area under ROC

Accuracy

$$\boxed{\frac{1}{m}\sum_{i=1}^{m}[a(x_i) = y_i]}$$

- Intuitive
- Have problems with unbalanced sets

Confusion matrix

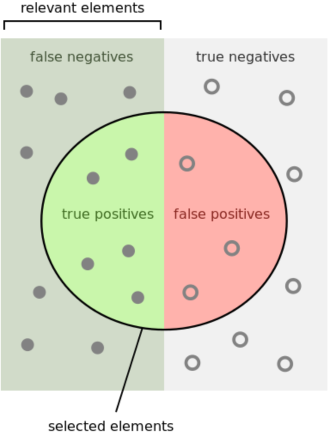
	$y = 1$	$y = 0$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = 0$	False Negative (FN)	True Negative (TN)

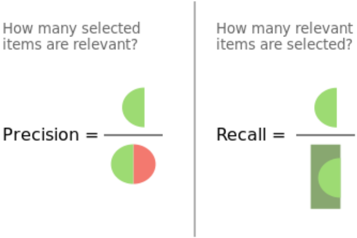
$$accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

Confusion matrix

	Spam (Predicted)	Non-Spam (Predicted)	Accuracy
Spam (Actual)	27	6	81.81
Non-Spam (Actual)	10	57	85.07
Overall Accuracy			83.44

Precision, recall. F-measure

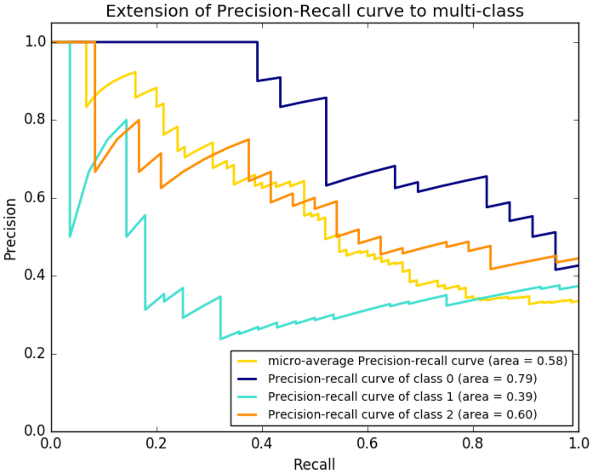




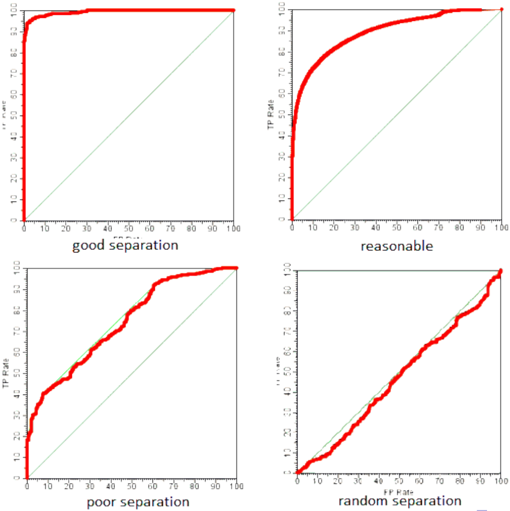
$$precision = \frac{TP}{TP+FP} \quad recall = \frac{TP}{TP+FN}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

PR - curve


Extension of Precision-Recall curve to multi-class
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
micro-average Precision-recall curve (area = 0.58)
Precision-recall curve of class 0 (area = 0.79)
Precision-recall curve of class 1 (area = 0.39)
Precision-recall curve of class 2 (area = 0.60)

ROC



good separation
reasonable
poor separation
random separation
TP Rate
FP Rate

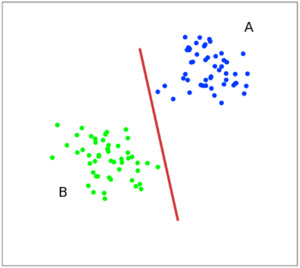
Metrics

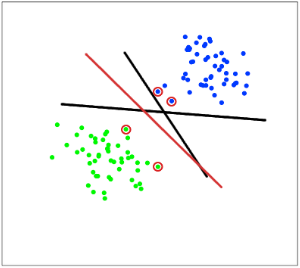
1. Accuracy - simple, but may be not objective
2. Precision, recall - more complex, objective
3. Confusion matrix - allows to see confusion between categories
4. F-measure - one metric for precision and recall
5. PR-curve - allows to choose the threshold
6. Area under PR-curve
7. Area under ROC

Support vector machines (SVM)

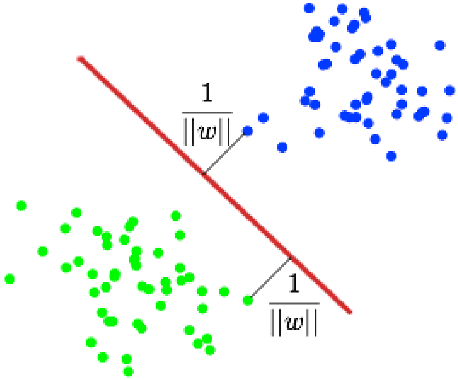
$$f(x) = w^T \cdot x + b$$

$$hinge\ loss = max(0, 1 - y_i \cdot f(x_i))$$

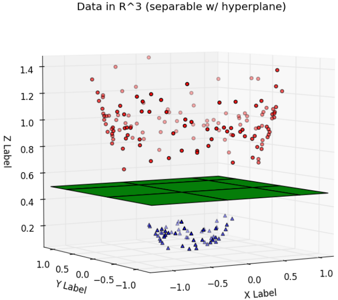


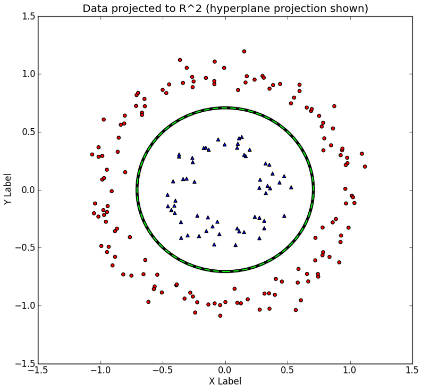


Support vector machines (SVM)



Support vector machines (SVM)





Support vector machines (SVM)

Kernel types:

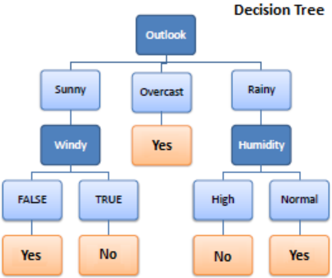
- Polynomial
- Radial basis function
- Gaussian radial basis function
- Sigmoid

Support vector machines (SVM)

- Find optimal solution
- Have problems with big datasets

Decision trees

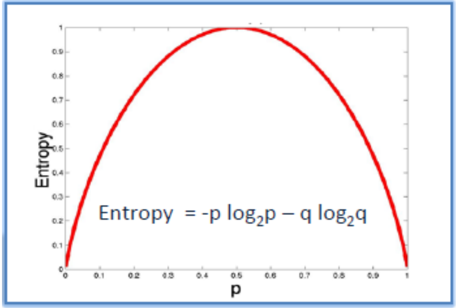
Predictors				Target
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Decision trees. ID3 algorithm

```
ID3 (Examples, Target_Attribute, Attributes)
    Create a root node for the tree
    If all examples are positive, Return the single-node tree Root, with label = +.
    If all examples are negative, Return the single-node tree Root, with label = -.
    If number of predicting attributes is empty, then Return the single node tree Root,
    with label = most common value of the target attribute in the examples.
    Otherwise Begin
        A ← The Attribute that best classifies examples.
        Decision Tree attribute for Root = A.
        For each possible value, vᵢ, of A,
            Add a new tree branch below Root, corresponding to the test A = vᵢ.
            Let Examples(vᵢ) be the subset of examples that have the value vᵢ for A
            If Examples(vᵢ) is empty
                Then below this new branch add a leaf node with label = most common target value in the examples
            Else below this new branch add the subtree ID3 (Examples(vᵢ), Target_Attribute, Attributes – {A})
    End
    Return Root
```

Decision trees. Entropy



Entropy = -0.5 $\log_2$0.5 – 0.5 $\log_2$0.5 = 1

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

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

Decision trees. Information gain

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

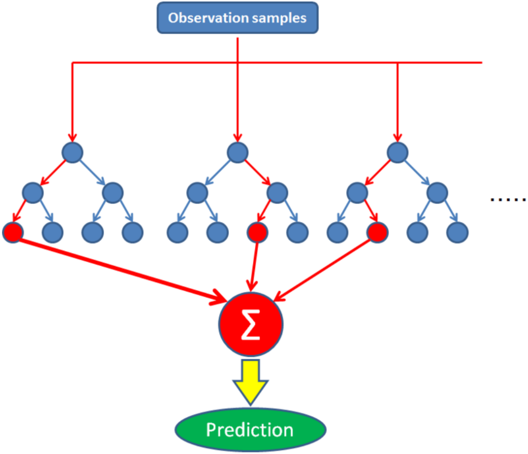
G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

= 0.940 – 0.693 = 0.247

Decision trees

- Interpreted
- Possible to apply to data with missing values

Random forest



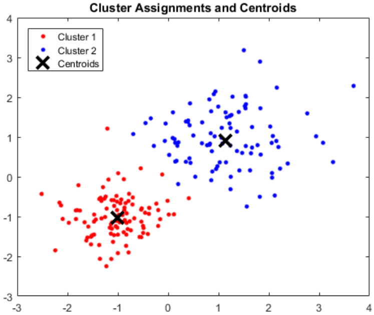
Random forest

- Good with big dimension data
- Easy to parallel and scale

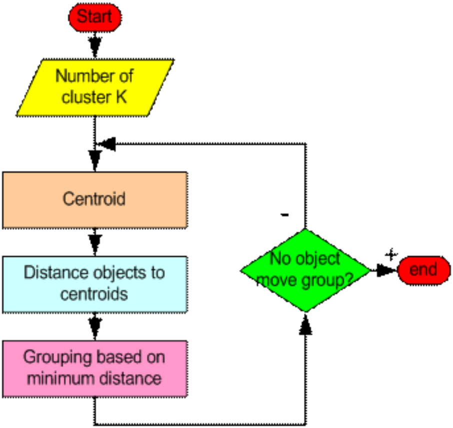
Cluster analysis. Types

- Hierarchical (Complete-linkage clustering)
- Centroid-based (K-means)
- Distributed-based (expectation–maximization (EM) algorithm)
- Density-based clustering (DB-SCAN)

Clusterisation. K-means



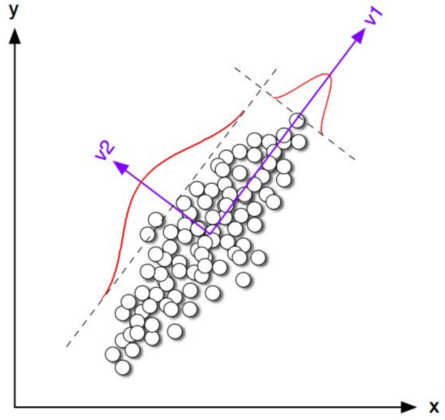
Clusterisation. K-means



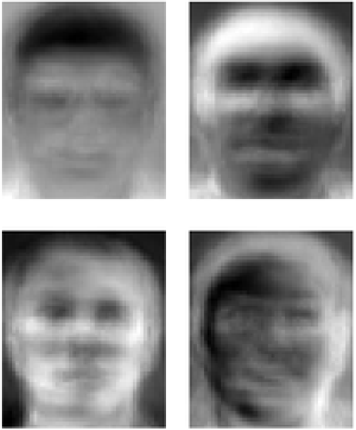
Clusterisation. K-means

- Simple
- Fast
- Need to know number of clusters (unknown optimal number)
- Don't achieve global minimum

Dimension reduction. PCA



Dimension reduction. PCA



Dimension reduction. PCA

- Works with big datasets
- Works poorly with nonlinear space