# Naïve Bayes

## Applications:

### Medical applications of Gaussian naïve bayes

R. Bhuvaneswari and K. Kalaiselvi, *Naive Bayesian Classification Approach in Healthcare Applications*. International Journal of Computer Science and Telecommunications, 2012. 3(1): p. 106-112

### Spam detection

https://opendatascience.com/blog/naive-bayes-and-spam-detection/

### Stock market predictions

I. S. L. Sarwani, N. Siva Nandhalahari, S. Sobha Sri, *Comparative Analysis of Id3 and Naïve Bayes Algorithm on Stock Market Prediction*. International Journal of Computer Applications (0975 – 8887) Volume 143 – No.2, June 2016

## Advantages:

### Fast algorithm

<https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algo_nb.htm#BABIIDDE>

### Works well with small datasets

http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/

## Disadvatages:

### Necessity to discretize features with real values to estimate probabilities

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (pp. 338-345). Montreal, Quebec: Morgan Kaufmann.

### Ok for classification: Assigns maximum probability to correct class. Because of independence assumptions, inaccurate probability values may lead to inaccurate results in the case of regression.

Eibe Frank, Leonard E. Trigg, Geoffrey Holmes, and Ian H. Witten. Naive Bayes for regression (technical note). Machine Learning, 41(1):5-25, 2000

### Cannot learn interaction between features because of independence assumptions

<http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/>

## How will it performs given what we know about the data?

## What we know about the data:

### 45222 instances

### 103 features

### Continuous and binary variables

This algorithm could be overperformed by others that perform bets with big datasets

# K-NN

## Industrial applications:

### Economic forecasting prediction

<http://www.academia.edu/4607757/Application_of_K-Nearest_Neighbor_KNN_Approach_for_Predicting_Economic_Events_Theoretical_Background>

### Face recognition

<https://www.quora.com/What-are-industry-applications-of-the-K-nearest-neighbor-algorithm?share=1>

## Advantages:

### Lazy learner: No training time, no need to actualize a model when with instances

### Adaptative behavior as it uses local information

### Works well with big datasets

<http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/>

### Ability to provide a range of values

<https://www.quora.com/What-are-industry-applications-of-the-K-nearest-neighbor-algorithm?share=1>

## Disadvantages:

Need to specify K

Need to determine the type of distance to use

Need to determine the type of “central tendency” to compute for the K chosen points (mean, median, mod, etc…)

Need to compute the distance with each training example to predict the output -> Can be slow with large datasets. Curse of dimensionality

Prone to local noise with small K

The density estimate diverge over the whole sample space (CSCE 666 Pattern Analysis | Ricardo Gutierrez Osuna | CSE@TAMU)

## How will it performs given what we know about the data?

## What we know about the data:

### 45222 instances

### 103 features

### Continuous and binary variables

KNN may be desirable when data is perceived as unstructured or when there is not much knowledge about it. Moreover, it may perform better than Naïve Bayes with big dataset.

# SVM

## Advantages:

### Robust to noise

<http://condor.depaul.edu/ntomuro/courses/578/notes/SVM-overview.pdf> (SVM overview by Noriko Tomuro Associate professor @ DePaul University)

### Highly accurate

<http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/>

### Less prone to overfitting

<http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/>

### No need for data linearly separable

## Disadvantages:

### Computationally expansive

<http://condor.depaul.edu/ntomuro/courses/578/notes/SVM-overview.pdf> (SVM overview by Noriko Tomuro Associate professor @ DePaul University)

### Can be sensitive to overfitting the hyperparameters of the model (regularization parameter and kernel parameter)

G. C. Cawley and N. L. C. Talbot, Over-fitting in model selection and subsequent selection bias in performance evaluation, Journal of Machine Learning Research, 2010. Research, vol. 11, pp. 2079-2107, July 2010.

### Long to tune

<http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/>

### Sensitive to noise

<https://eric.univ-lyon2.fr/~ricco/cours/slides/en/svm.pdf>

### Difficulties with large datasets

<https://eric.univ-lyon2.fr/~ricco/cours/slides/en/svm.pdf>

### Robust when few observations compared to number of variables

<https://eric.univ-lyon2.fr/~ricco/cours/slides/en/svm.pdf>

## Applications:

### SVM outperforms backpropagation neural network for financial time series forecasting

TAY, Francis E. H. and Lijuan CAO, 2001. [Application of support vector machines in financial time series forecasting](http://dx.doi.org/10.1016/S0305-0483%2801%2900026-3), *Omega: The International Journal of Management Science*, Volume 29, Issue 4, August 2001, Pages 309-317. [[Cited by 67](http://scholar.google.com/scholar?hl=en&lr=&ie=UTF-8&cites=195295496737709481)]

CHEN, Wun-Hwa and Jen-Ying SHIH, 2006. [A study of Taiwan's issuer credit rating systems using support vector machines](http://svms.org/finance/ChenShih2006.pdf), *Expert Systems with Applications*, Volume 30, Issue 3, April 2006, Pages 427-435.

CHEN, Wun-Hua, Jen-Ying SHIH and Soushan WU, 2006. [Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock markets](http://svms.org/finance/ChenShihWu2006.pdf), *International Journal of Electronic Finance*, Volume, Issue 1, pages 49-67.

### Industrial application for mechanical faults diagnostic

[Lane Maria Rabelo Baccarini](http://www.sciencedirect.com/science/article/pii/S0957417410013801), , [Valceres Vieira Rocha e Silva](http://www.sciencedirect.com/science/article/pii/S0957417410013801), [Benjamim Rodrigues de Menezes](http://www.sciencedirect.com/science/article/pii/S0957417410013801), [Walmir Matos Caminhas](http://www.sciencedirect.com/science/article/pii/S0957417410013801), *SVM practical industrial application for mechanical faults diagnostic*. [Expert Systems with Applications](http://www.sciencedirect.com/science/journal/09574174), [Volume 38, Issue 6](http://www.sciencedirect.com/science/journal/09574174/38/6), June 2011, Pages 6980–6984

# Decision Trees

## Industrial applications:

### Risk management

<http://projectrisk.com/white_papers/Use_Decision_Trees_to_Make_Important_Project_Decisions.pdf>

### Quality engineering in a semiconductor industry

<https://www.researchgate.net/profile/A_Kusiak/publication/3424896_Decomposition_in_data_mining_An_industrial_case_study/links/550c25850cf2528164db7641/Decomposition-in-data-mining-An-industrial-case-study.pdf>

## Disadvantages:

### Exponential storage growth while number of features is increasing

### Small training sets result in high classification error metrics

### Have to rebuilt the tree when learning new instances

### Prone to overfitting… but ensemble methods can help to solve this issue

<http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/>

## Advantages:

### Generalization without much calculations

### Can handle continuous and discrete features

### Data don’t have to be linearly separable

# Ada Boost

## Industrial Applications:

### Economic predictions

[Soo Y. Kim](http://www.sciencedirect.com/science/article/pii/S0264999313004318), [Arun Upneja](http://www.sciencedirect.com/science/article/pii/S0264999313004318), *Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models*. [Economic Modelling](http://www.sciencedirect.com/science/journal/02649993), [Volume 36](http://www.sciencedirect.com/science/journal/02649993/36/supp/C), January 2014, Pages 354–362

### Solder Joint inspection of chip component

Xie Hongwei, Zhang Xianmin, Kuang Yongcong, [Ouyang Gaofei](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Ouyang%20Gaofei.QT.&newsearch=true), *Solder Joint Inspection Method for Chip Component Using Improved AdaBoost and Decision Tree*. [IEEE Transactions on Components, Packaging and Manufacturing Technology](http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=5503870) ( Volume: 1, [Issue: 12](http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=6105480), Dec. 2011 )

### Visual recognition

<https://core.ac.uk/download/pdf/21751148.pdf>

### Biology, speech processing

<http://www.nickgillian.com/wiki/pmwiki.php?n=GRT.AdaBoost>

## Advantages:

### Few parameters to tune

<http://www.nickgillian.com/wiki/pmwiki.php?n=GRT.AdaBoost>

### Good generalization

<http://www.robots.ox.ac.uk/~az/lectures/cv/adaboost_matas.pdf>

### Fast

<http://cseweb.ucsd.edu/~yfreund/papers/IntroToBoosting.pdf>

## Disadvantages:

##### Sensitive to noisy data and outliers:

<http://www.nickgillian.com/wiki/pmwiki.php?n=GRT.AdaBoost>

## Why this classifier would be a good fit for the data?

The data isn’t signal from captors. Categorical data. => Less likely to have noise

Skewed variables have been preprocessed => No outliers