Data Mining

John Samuel
CPE Lyon

Year: 2019-2020
Email: john(dot)samuel(at)cpe(dot)fr

Goals

- Understanding Patterns
- Data mining tasks
- Algorithms for data mining
- Feature Selection

1. Patterns




Patterns in Nature

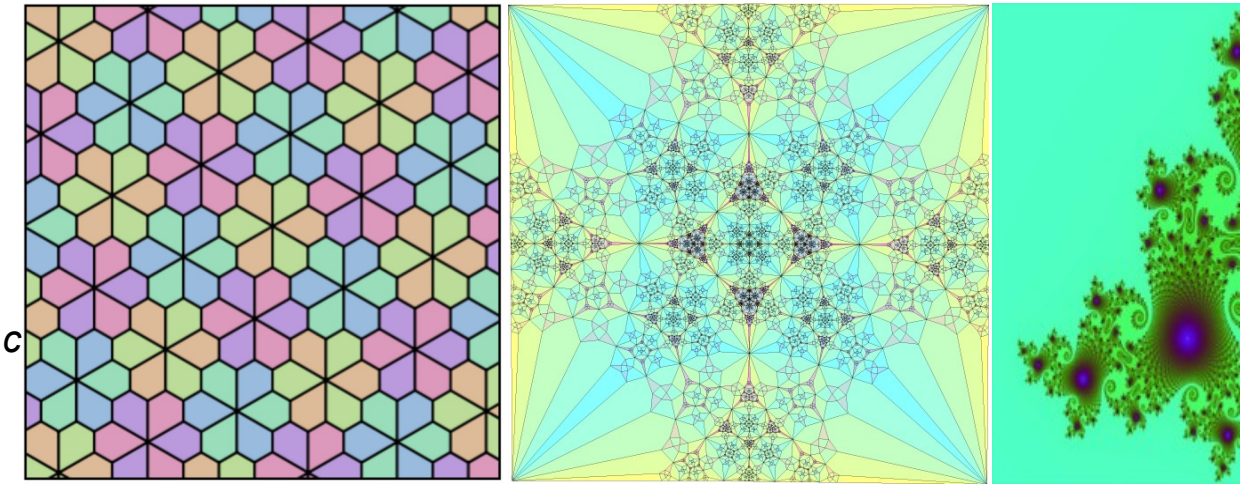
- Symmetry
- Trees, Fractals
- Spirals
- Chaos
- Waves
- Bubbles, Foam
- Tesselations
- Cracks
- Spots, stripes

Patterns by Humans

- Buildings (Symmetry)
- Cities
- Virtual environments (e.g., video games)
- Human artifacts

Pattern creation

- Repitition
- Fractals
  - Julia set: $f(z) = z^2 + c$



Synonyms

- Pattern recognition
- Knowledge discovery in databases
- Data mining
- Machine learning

Pattern Recognition

- Goal is to detect patterns and regularities in data
- Approaches
  1. **Supervised learning**: Availability of labeled training data
  2. **Unsupervised learning**: No labeled training data available
  3. **Semi-supervised learning**: Small set of labeled training data and a large amount of unlabeled data

Formalization

- **Euclidean vector**: geometric object with magnitude and direction
- **Vector space**: collection of vectors that can be added together and multiplied by numbers.
- **Feature vector**: n-dimensional vector
- **Feature space**: Vector space associated with the vectors

Examples: Features

- **Images**: pixel values.
- **Texts**: Frequency of occurence of textual phrases.

Formalization

- **Feature construction**[1]: construction of new features from already available features
- **Feature construction operators**
  - Equality operators, arithmetic operators, array operators (min, max, average etc.)...

Example

- Let **Year of Birth** and **Year of Death** be two existing features.
- A new feature called **Age** = Year of Birth - Year of Death

1. https://en.wikipedia.org/wiki/Feature_vector

Formalization: Supervised learning

- Let $\boldsymbol{N}$ be the number of training examples
- Let $\boldsymbol{X}$ be the input feature space
- Let $\boldsymbol{Y}$ be the output feature space (of labels)
- Let $\{(x_1, y_1),...,(x_{\boldsymbol{N}}, y_{\boldsymbol{N}})\}$ be the $\boldsymbol{N}$ training examples, where
  - x_i is the feature vector of i^{th} training example.
  - y_i is its label.
- The goal of supervised learning algorithm is to find $g: \boldsymbol{X} \rightarrow \boldsymbol{Y}$, where
  - g is one of the functions from the set of possible functions G (hypotheses space)
- **Scoring function** $\boldsymbol{F}$ denote the space of scoring functions, where
  - $f: \boldsymbol{X} \times \boldsymbol{Y} \rightarrow \boldsymbol{R}$ such that g returns the highest scoring function.

Formalization: Unsupervised learning

- Let $\boldsymbol{X}$ be the input feature space
- Let $\boldsymbol{Y}$ be the output feature space (of labels)
- The goal of unsupervised learning algorithm is to
  - find mapping $\boldsymbol{X} \rightarrow \boldsymbol{Y}$

Formalization: Semi-supervised learning

- Let $\boldsymbol{X}$ be the input feature space
- Let $\boldsymbol{Y}$ be the output feature space (of labels)
- Let $\{(x_1, y_1),...,(x_l, y_l)\}$ be the $\boldsymbol{l}$ be the set of labeled training examples
- Let $\{x_{l+1},...,x_{l+u}\}$ be the $\boldsymbol{u}$ be the set of unlabeled feature vectors of $\boldsymbol{X}$.
- The goal of semi-supervised learning algorithm is to do
  - **Transductive learning**, i.e., find correct labels for $\{x_{l+1},...,x_{l+u}\}$. OR
  - **Inductive learning**, i.e., find correct mapping $\boldsymbol{X} \rightarrow \boldsymbol{Y}$

Tasks in Data Mining

1. Classification
2. Clustering
3. Regression
4. Sequence Labeling
5. Association Rules
6. Anomaly Detection
7. Summarization

2.1. Classification

- Generalizing known structure to apply to new data
- Identifying the set of categories to which an object belongs
- Binary vs. Multiclass classification

Applications

- Spam vs Non-spam
- Document classification
- Handwriting recognition
- Speech Recognition
- Internet Search Engines

Formal definition

- Let $\boldsymbol{X}$ be the input feature space
- Let $\boldsymbol{Y}$ be the output feature space (of labels)
- The goal of classification algorithm (or classifier) is to find $\{(x_1, y_1),...,(x_l, y_k)\}$, i.e., assigning a known label to every input feature vector, where
  - $x_i \in \boldsymbol{X}$
  - $y_i \in \boldsymbol{Y}$
  - $|\boldsymbol{X}| = l$
  - $|\boldsymbol{Y}| = k$
  - l >= k

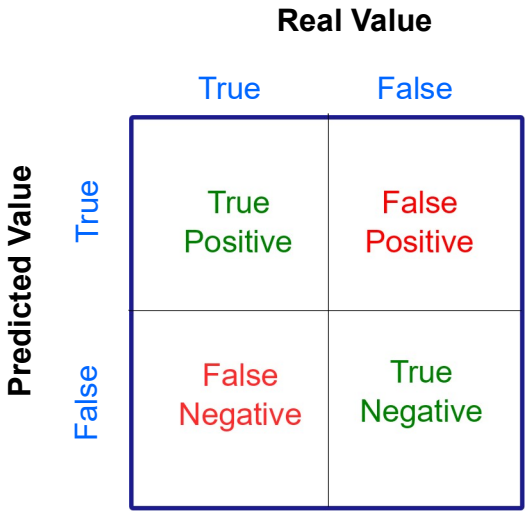
Classifiers

- Classifying Algorithm
- Two types of classifiers:
  - **Binary classifiers** assigning an object to any of two classes
  - **Multiclass classifiers** assigning an object to one of several classes

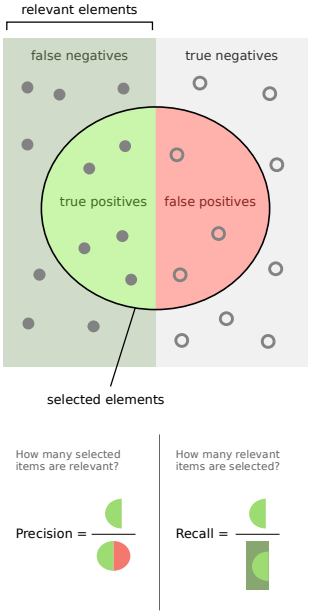
Linear Classifiers

- A linear function assigning a score to each possible category by combining the feature vector of an instance with a vector of weights, using a dot product.
- Formalization:
  - Let $\boldsymbol{X}$ be the input feature space and $\boldsymbol{x}_i \in \boldsymbol{X}$
  - Let $\boldsymbol{\beta}_k$ be vector of weights for category k
  - $score(\boldsymbol{x}_i, k) = \boldsymbol{x}_i.\boldsymbol{\beta}_k$, score for assigning category k to instance $\boldsymbol{x}_i$. The category that gives the highest score is assigned as the category of the instance.

2.2. Classifiers



2.2. Classifiers



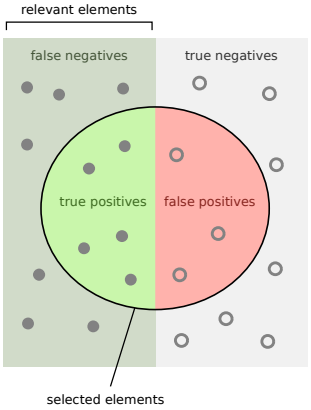
2.2. Classifiers

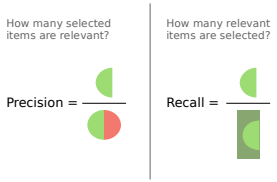
Let

- *tp*: number of true postives
- *fp*: number of false postives
- *fn*: number of false negatives

Then

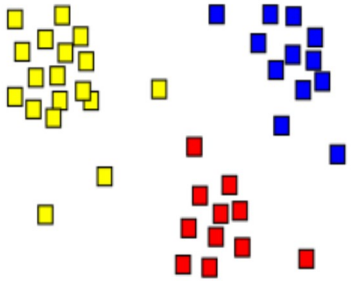
- Precision *p = tp / (tp + fp)*
- Recall *r = tp / (tp + fn)*
- F1-score *f1 = 2 * ((p * r) / (p + r))*





2.2. Clustering

- Discovering groups and structures in the data without using known structures in the data
- Objects in a cluster are more similar to each other than the objects in the other cluster



Applications

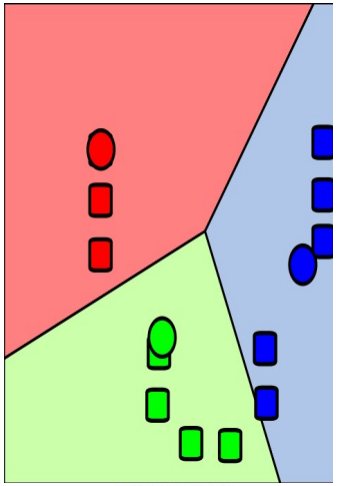
- Social network analysis
- Image segmentation
- Recommender systems
- Grouping of shopping items

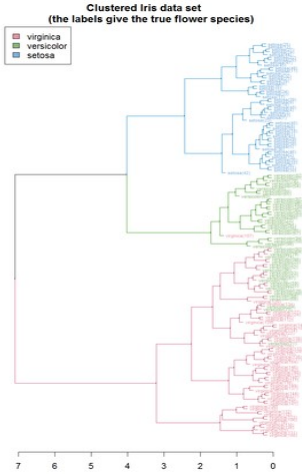
Formal definition

- Let $\boldsymbol{X}$ be the input feature space
- The goal of clustering is to find k subsets of $\boldsymbol{X}$, in such a way that
  - $C_1 .. \cup .. C_k \cup C_{outliers} = \boldsymbol{X}$ and
  - $C_i \cap C_j = \phi,\ i \neq j;\ 1 < i,j < k$
  - $C_{outliers}$ may consist of outlier instances (data anomaly)

Cluster models

- **Centroid models**: cluster represented by a single mean vector
- **Connectivity models**: distance connectivity
- Distribution models: clusters modeled using statistical distributions
- Density models: clusters as connected dense regions in the data space
- Subspace models
- Group models
- Graph-based models
- Neural models



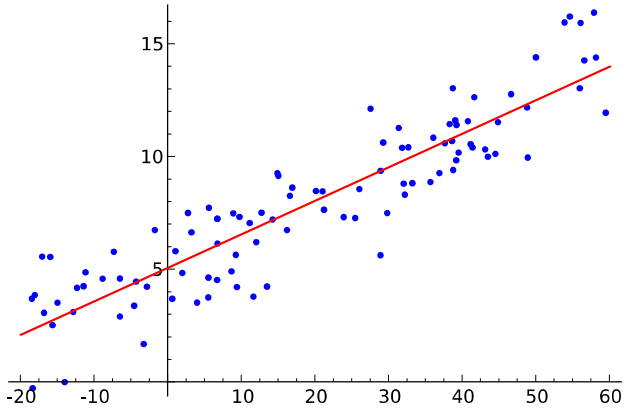


2.3. Regression

- Finding a function which models the data
- Assigns a real-valued output to each input
- Estimating the relationships among variables
- Relationship between a dependent variable ('criterion variable') and one or more independent variables ('predictors').

Applications

- Prediction
- Forecasting
- Machine learning
- Finance



Formal definition

- A function that maps a data item to a prediction variable
- Let $\boldsymbol{X}$ be the independent variables
- Let $\boldsymbol{Y}$ be the dependent variables
- Let $\boldsymbol{\beta}$ be the unknown parameters (scalar or vector)
- The goal of regression model is to approximate $\boldsymbol{Y}$ with $\boldsymbol{X},\boldsymbol{\beta}$, i.e.,
  - $\boldsymbol{Y} \cong f(\boldsymbol{X},\boldsymbol{\beta})$

Linear regression

- straight line: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ OR
- parabola: $y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2 + \varepsilon_i$

Linear regression

- straight line: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ OR
- $\hat{y}_i = \beta_0 + \beta_{1i}$ OR
- Residual: $e_i = \hat{y}_i - y_i$
- Sum of squared residuals, SSE = $\Sigma\ e_i$, where $1 < i < n$
- The goal is to minimize SSE

2.4. Sequence Labeling

- Assigning a class to each member of a sequence of values

Applications

- Part of speech tagging
- Linguistic translation
- Video analysis
- Handwriting recognition
- Information extraction

Formal definition

- Let $\boldsymbol{X}$ be the input feature space
- Let $\boldsymbol{Y}$ be the output feature space (of labels)
- Let $\langle x_1,...,x_T \rangle$ be a sequence of length T.
- The goal of sequence labeling is to generate a corresponding sequnce
  - $\langle y_1,...,y_T \rangle$ of labels
  - $x_i \in \boldsymbol{X}$
  - $y_j \in \boldsymbol{Y}$

Association Rules

- Searches for relationships between variables

Applications

- Web usage mining
- Intrusion detection
- Affinity analysis

Formal definition

- Let $\boldsymbol{I}$ be a set of n binary attributes called items
- Let $\boldsymbol{T}$ be a set of m transactions called database
- Let $\boldsymbol{I} = \{(i_1,...,i_n)\}$ and $\boldsymbol{T} = \{(t_1,...,t_m)\}$
- The goal of association rule learning is to find
  - $\boldsymbol{X} \Rightarrow \boldsymbol{Y}$, where $\boldsymbol{X} \Rightarrow \boldsymbol{Y} \subseteq \boldsymbol{I}$
  - $\boldsymbol{X}$ is the antecedent
  - $\boldsymbol{Y}$ is the consequent

Formal definition

- Support: how frequently an itemset appears in the database
  - $supp(\boldsymbol{X}) = |t \in \boldsymbol{T}; \boldsymbol{X} \subseteq t| \,/\, |\boldsymbol{T}|$
- Confidence: how frequently the rule has been found to be true.
  - $conf(\boldsymbol{X} \Rightarrow \boldsymbol{Y}) = supp(\boldsymbol{X} \cup \boldsymbol{Y})/supp(\boldsymbol{X})$
- Lift: the ratio of the observed support to that of the expected if X and Y were independent
  - $lift(\boldsymbol{X} \Rightarrow \boldsymbol{Y}) = supp(\boldsymbol{X} \cup \boldsymbol{Y})/(supp(\boldsymbol{X}) \times supp(\boldsymbol{Y}))$

Example

- {**bread**, **butter**} ⇒ {**milk**}

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

2.6. Anomaly Detection

- Identification of unusual data records
- Approaches
  1. Unsupervised anomaly detection
  2. Supervised anomaly detection
  3. Semi-supervised anomaly detection

Applications

- Intrusion detection
- Fraud detection
- Remove anomalous data
- System health monitoring
- Event detection in sensor networks
- Misuse detection

Characteristics

- Unexpected bursts

Formalization

- Let $\boldsymbol{Y}$ be a set of measurements
- Let $P_{\boldsymbol{Y}}(y)$ be a statistical model for the distribution of $\boldsymbol{Y}$ under 'normal' conditions.
- Let $\boldsymbol{T}$ be a user-defined threshold.
- A measurement x is an outlier if $P_{\boldsymbol{Y}}(x) < \boldsymbol{T}$

2.7. Summarization

- Providing a more compact representation of the data set
- Report Generation

Applications

- Keyphrase extraction
- Document summarization
- Search engines
- Image summarization
- Video summarization: Finding important events from videos

Formalization: Multidocument summarization

- Let {$\boldsymbol{D} = D_1, ..., D_k$} be a document collection of k documents
- A Document {$D = t_1, ..., t_m$} consists of m textual units (words, sentences, paragraphs etc.)
- Let {$\boldsymbol{D} = t_1, ..., t_n$} be the complete set of all textual units from all documents, where
  - $t_i \in \boldsymbol{D}$, if and only if $\exists\, D_j$ such that $t_i \in D_j$
- $S \subseteq \boldsymbol{D}$ constitutes a summary
- Two scoring functions
  - *Rel(i)*: relevance of textual unit *i* in the summary
  - *Red(i,j)*: Redundancy between two textual units t_i, t_j

Formalization: Multidocument summarization

- Scoring for a summary S
  - *s(S)* score of summary S
  - *l(i)* is the length of the textual unit i
  - *K* is the fixed maximum length of the summary

$$S = \arg\max_{S \subseteq \boldsymbol{D}} \; s(S)$$

$$= \arg\max_{S \subseteq \boldsymbol{D}} \sum_{t_i \in S} Rel(i) \; - \sum_{t_i, t_j \in S,\, i<j} Red(i,j)$$

$$\text{such that} \sum_{t_i \in S} l(i) \leq K$$

2.7. Summarization

- Finding a subset from the entire subset
- Approaches
  1. **Extraction**: Selecting a subset of existing words, phrases, or sentences in the original text without any modification
  2. **Abstraction**: Build an internal semantic representation and then use natural language generation techniques

Extractive summarization

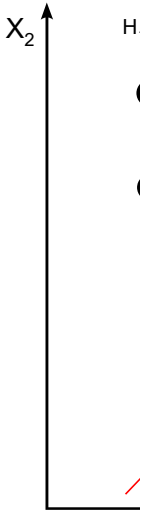
- Approaches
  1. **Generic summarization**: Obtaining a generic summary
  2. **Query relevant summarization**: Summary relevant to a query

3. Algorithms

1. Support Vector Machines (SVM)
2. Stochastic Gradient Descent (SGD)
3. Nearest-Neighbours
4. Naive Bayes
5. Decision Trees
6. Ensemble Methods (Random Forest)

Introduction

- Supervised learning approach
- Binary classification algorithm
- Constructs a hyperplane ensuring the maximum separation between two classes

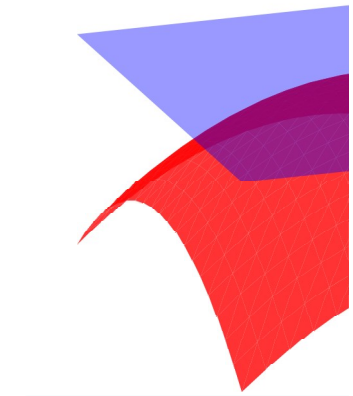


Hyperplane

- Hyperplane of n-dimensional space is a subspace of dimension *n-1*
- Examples
  - Hyperplane of a 2-dimensional space is 1-dimensional line
  - Hyperplane of a 3-dimensional space is 2-dimensional plane

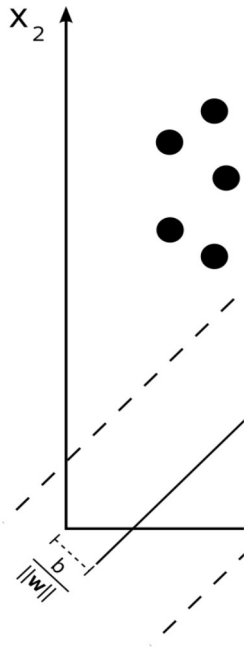
Formal definition

- The goal of a SVM is to estimate a function $f: R^N \times \{+1,-1\}$, i.e.,
  - If $x_1,...,x_l \in R^N$ are the N input data points,
  - the goal is to find $(x_1,y_1),...,(x_l,y_l) \in R^N \times \{+1,-1\}$
- Any hyperplane can be written by the equation using set of input points $\boldsymbol{x}$
  - $\boldsymbol{w}.\boldsymbol{x} - b = 0$, where
  - $\boldsymbol{w} \in R^N$, a normal vector to the plane
  - $b \in R$
- A decision function is given by $f(x) = sign(\boldsymbol{w}.\boldsymbol{x} - b)$



Formal definition

- If the training data are linearly separable, two hyperplanes can be selected
- They separate the two classes of data,
  so that distance between them is as large as possible.
- The hyperplanes can be given by the equations
  - $\boldsymbol{w}.\boldsymbol{x} - b = 1$
  - $\boldsymbol{w}.\boldsymbol{x} - b = -1$
- The distance between the two hyperplanes can be given by $2/||\boldsymbol{w}||$
- Region between these two hyperplanes is called margin.
- Maximum-margin hyperplane is the hyperplane
  that lies halfway between them.



Formal definition

- In order to prevent data points from falling into the margin, following constraints are added
  - $\boldsymbol{w}.\boldsymbol{x}_i - b >= 1$, if $y_i = 1$
  - $\boldsymbol{w}.\boldsymbol{x}_i - b <= -1$, if $y_i = -1$
- $y_i(\boldsymbol{w}.\boldsymbol{x}_i - b) >= 1$ for 1<= i <= n
- The goal is to minimize ||**w**|| subject to $y_i(\boldsymbol{w}.\boldsymbol{x}_i - b) >= 1$ for 1<= i <= n
- Solving for both ***w*** and *b* gives our classifier $f(x) = sign(\boldsymbol{w}.\boldsymbol{x} - b)$
- Max-margin hyperplane is completely determined by the points that lie nearest to it, called the **support vectors**

Data mining tasks

- Classification (Multi-class classification)
- Regression
- Anomaly detection

Applications

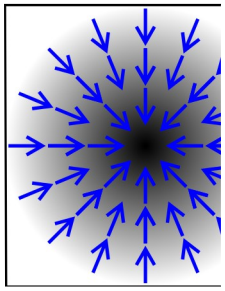
- Text and hypertext categorization
- Image classification
- Handwriting recognition

3.2. Stochastic Gradient Descent (SGD)

- A stochastic approximation of the gradient descent optimization
- Iterative method for minimizing an objective function that is written as a sum of differentiable functions.
- Finds minima or maxima by iteration

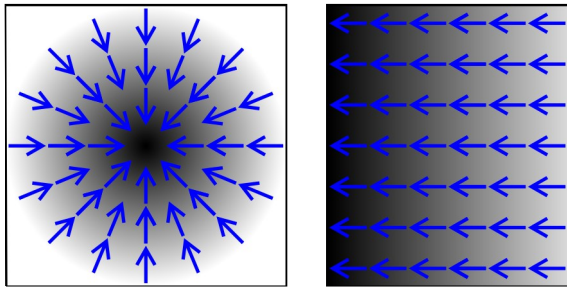
Gradient

- Multi-variable generalization of the derivative.
- Gives slope of the tangent of the graph of a function
- Gradient points in the direction of the greatest rate of increase of a function
- Magnitude of gradient is the slope of the graph in that direction



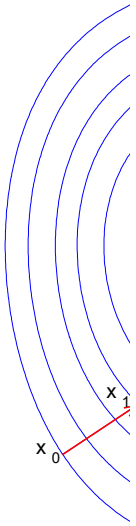
Gradient vs Derivative

- Derivatives defined on functions of single variable
- Gradient defined on functions of multiple variables
- Gradient is a vector-valued function (range is a vector)
- Derivative is a scalar-valued function



Gradient descent

- First-order iterative optimization algorithm for finding the minimum of a function.
- Finding a local minima involves taking steps proportional to the negative of the gradient of the function at the current point.



Standard gradient descent method

- Let's take the problem of minimizing an objective function
  - $Q(w) = 1/n (\Sigma Q_i(w)), 1<=i<n$
  - Summand function Q_i associated with i^{th} observation in the data set.
- $w = w - \eta . \nabla Q(w)$

Iterative method

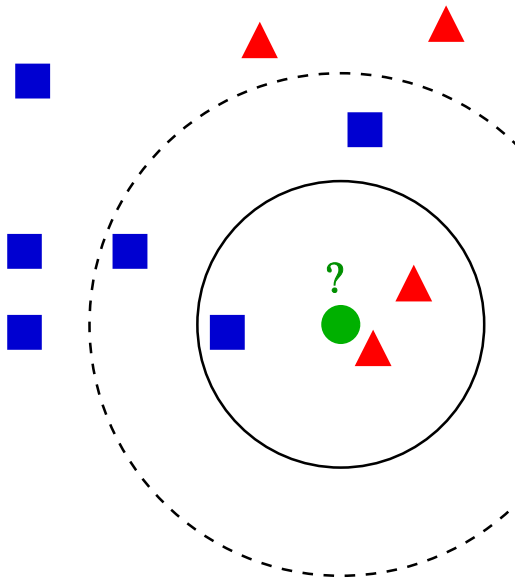
- Choose an initial vector of parameters and learning rate η.
- Repeat until an approximate minimum is obtained:
  - Randomly shuffle examples in the training set.
  - $w = w - \eta . \nabla Q_i(w)$, for i=1...n

Applications

- Classification
- Regression

k-nearest neighbors algorithm

- k-NN classification: output is a class membership
  (object is classified by a majority vote of its neighbors.)
- k-NN regression: output is the property value for the object
  (average values of its k nearest neighbors)



Applications

- Regression
- Anomaly detection

3.4. Naive Bayes classifiers

- Collection of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumption between the features.

Applications

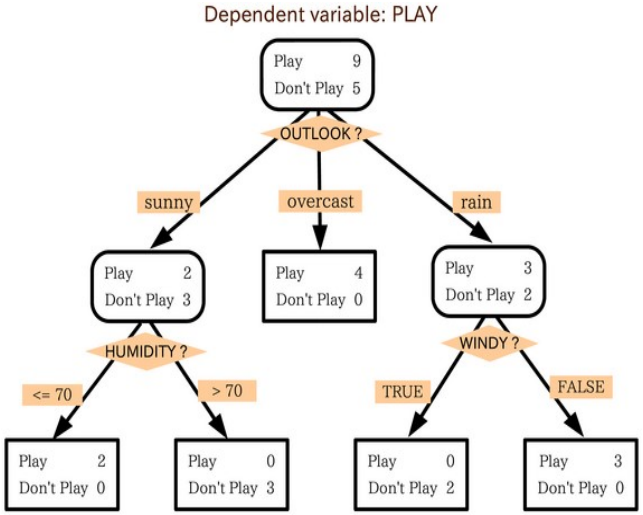
- Document classification (spam/non-spam)

Bayes' Theorem

- If A and B are events.
- P(A), P(B) are probabilities of observing A and B independently of each other..
- P(A|B) is conditional probability, the likelihood of event A occurring given that B is true
- P(B|A) is conditional probability, the likelihood of event B occurring given that A is true
- P(B) ≠ 0
- $P(A|B) = (P(B|A).P(A))/P(B)$

3.5. Decision Trees

- Decision support tool
- Tree-like model of decisions and their possible consequences



Applications

- Classification
- Regression
- Decision Analysis: identifying strategies to reach a goal
- Operations Research

Defintion

- Collection of multiple learning algorithms to obtain better predictive performance than could be obtained from one of the constituting algorithms alone.
- Random forests are obtained by building multiple decision trees at training time

3.6. Ensemble Methods (Random Forest)

- Multiclass classification
- Multilabel classification (the problem of assigning one or more label to each instance. There is no limit on the number of classes an instance can be assigned to.)
- Regression
- Anomaly detection

Definition

- Process of selecting a subset of relevant features
- Used in domains with large number of features and comparatively few sample points

Applications

- Analysis of written texts
- Analysis of DNA microarray data

Formal defintion[8]

- Let X be the original set of n features, i.e., $|X| = n$
- Let w_i be the weight assigned to feature $x_i \in X$
- Binary feature selection assigns binary weights whereas continuous feature selection assigns weights preserving the order of its relevance.
- Let $J(X')$ be an evaluation measure, defined as $J: X' \subseteq X \rightarrow R$
- Feature selection problem may be defined in three following ways
  1. $|X'| = m < n$. Find $X' \subset X$ such that $J(X')$ is maximum
  2. Choose J_0, Find $X' \subseteq X$, such that $J(X') >= J_0$
  3. Find a compromise among minimizing $|X'|$ and maximizing $J(X')$

Research articles

1. From data mining to knowledge discovery in databases, Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, AI Magazine Volume 17 Number 3 (1996)
2. Survey of Clustering Data Mining Techniques, Pavel Berkhin
3. Mining association rules between sets of items in large databases, Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD 1993. p. 207.
4. Comparisons of Sequence Labeling Algorithms and Extensions, Nguyen, Nam, and Yunsong Guo. Proceedings of the 24th international conference on Machine learning. ACM, 2007.

Research articles

5. An Analysis of Active Learning Strategies for Sequence Labeling Tasks, Settles, Burr, and Mark Craven. Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2008.
6. Anomaly detection in crowded scenes, Mahadevan; Vijay et al. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010
7. A Study of Global Inference Algorithms in Multi-Document Summarization. McDonald, Ryan. European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2007.
8. Feature selection algorithms: A survey and experimental evaluation., Molina, Luis Carlos, Lluís Belanche, and Àngela Nebot. Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE, 2002.
9. Support vector machines, Hearst, Marti A., et al. IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

Online resources

- Patterns in Nature
- Data Mining
- Statistical classification
- Regression analysis
- Cluster analysis
- Association rule learning
- Anomaly detection

- Sequence labeling
- Automatic summarization
- Pattern recognition
- Scikit-learn

Online resources

- Support Vector Machines
- Decision tree learning
- Stochastic gradient descent

References

Colors

- Color Tool - Material Design

Images

- Wikimedia Commons