Classification

Computer Vision

Ing. Ivan Gruber Ph.D.

Department of Cybernetics
Faculty of Applied Sciences
University of West Bohemia

ESF projekt Západočeské univerzity v Plzni
reg. č. CZ.02.2.69/0.0/0.0/16_015/0002287

Content

- Patter Recognition
  - General Knowledge
  - Supervised vs Unsupervised learning
- Clustering
  - General knowledge
  - K-Means
  - Agglomerative and Divisive clustering
  - Examples of Usage
- Supervised learning
  - General Knowledge
  - k-NN
  - Naive Bayes
  - SVM
  - Examples of Usage

Supervised vs Unsupervised learning

Supervised

- Data: (x, y)
- Goal: Learn a function to map x→y
- Examples: Classification, regression, object detection, semantic segmentation

Unsupervised

- Data: x
- Goal: Learn some underlying hidden structure of the data
- Examples: Clustering, dimensionality-reduction, feature learning, density estimation, etc

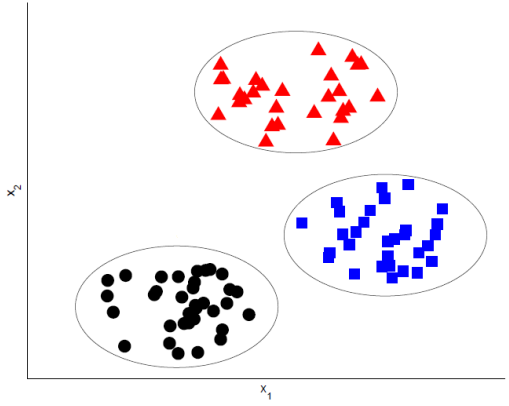
EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
DEPARTMENT OF CYBERNETICS

Classification

- Problem of identifying to which of a set of classes a new observation belongs
- Classifying based on **feature vectors**
  - n-dimensional vector describing attributes of the classified object/event
  - the set of all possible vectors forms n-dimensional feature space
- the task of binary classifier is to divide the feature space into two parts so that (ideally) all vectors from one class lie on one part of the space and vice versa
- a hyperplane is used as a solution of this problem (generally)

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
DEPARTMENT OF CYBERNETICS

- c_i is the i^{th} class, $X \in R^n$ is the space of all classes
- $c_i \cap c_j = \emptyset$



Feature vector

- the most important part is to describe the object based on the features that distinguish one from another
- the objects from one class should be similar in the feature space and different from objects of other classes
- Cosine similarity

$$s_{cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{||\mathbf{x}|| \, ||\mathbf{y}||} \tag{1}$$

- Distance measure

$$d_p(\mathbf{x}, \mathbf{y}) = (\sum_{i=1}^{I} \omega_i \, |\mathbf{x}_i - \mathbf{y}_i|^p)^{1/p} \tag{2}$$

- l_1-norm = Manhattan distance: $\omega = 1$ and $p = 1$
- l_2-norm = Euclidean distance: $\omega = 1$ and $p = 2$
- l_∞-norm: $d_\infty(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq I} |\mathbf{x}_i - \mathbf{y}_i|$

Vector and class comparisson

- the vector $\mathbf{x}$ is classified based on either all the vectors forming the class, or just on subset of vectors which are in some way representative
- Maximal similarity function:

$$\rho_{\max}(\mathbf{x}, c) = \max_{\mathbf{y} \in c} s(\mathbf{x}, \mathbf{y}) \tag{3}$$

- Minimal distance function:

$$\rho_{\min}(\mathbf{x}, c) = \min_{\mathbf{y} \in c} d(\mathbf{x}, \mathbf{y}). \tag{4}$$

- Average similarity function:

$$\rho_{\text{avg}}(\mathbf{x}, c) = \frac{1}{n_c} \sum_{\mathbf{y} \in c} d(\mathbf{x}, \mathbf{y}). \tag{5}$$

Clustering

- Unsupervised learning
- No data about correct class
- Goal: Divide patterns into K classes so that optimality criterion is minimal
- The most important methods:
  - K-Means (MacQueen algorithm - 1967)
  - Mean-Shift
  - Hierarchical clustering

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
DEPARTMENT OF CYBERNETICS

K-Means

- Most well-known clustering algorithm
- Algorithm:
  1. Place K points into the feature space. These points represent initial class centroids
  2. Assign each data point to the class that has the closest centroid (euclidean distance)
  3. Compute new centroids for the clusters by taking the average of the all data points that belong to each cluster
  4. If there is no change to the centroids, end. Else, go to 2.
- Advantages: Fast, linear complexity, easy to implement
- Disadvantages: Arbitrary choice of K (Elbow method), randomness, can handle only spherical shapes of clusters
- Modifications: Usage of closest point instead of 'real' centroid
- Example

Mean-shift

- A method for non-parametric probability density estimation
- Arbitrary choice of the size of kernel (size of sliding window)
- Algorithm:
  1. Randomly initialize centroids (grid)
  2. Assign all points within sliding window into the cluster
  3. Calculate locations of new centroid
  4. If there is no change to the centroids, end. Else, go to 2.
- Advantages: Does not make any assumption about probability distribution, non-convex clusters, robust to outliers
- Disadvantages: Problems with high dimensional data
- Example-1, Example-2

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Hierarchical clustering

- Top-down (Divisive) vs bottom-up (Agglomerative)
- Algorithm (Agglomerative):
  1. Assign each point into a separate cluster
  2. Combine two closest (most similar) clusters into new one
  3. Repeat until all the data are in one cluster
  4. The final result is chosen based on the length of life of clusters, or based on the maximum distance threshold
- Advantages: Does not require to specify the number of clusters, robust to choice of distance metric
- Disadvantages: Huge computational complexity
- Example

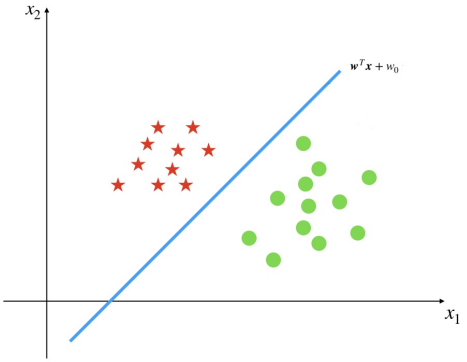
EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Supervised Learning

- Labeled data available (correct class for each pattern)
- During the training, patterns are presented to the classifier
- Classifier tries to predict correct class = classifier is trying to find correct **decision boundary parameters**
- Using **Loss function**, loss is calculated
- Correst parameters are found using **optimization method**

Decision boundary

- = hypersurface that partitions the underlying vector space into two sets (one for each class)
- The classification is then reduced to the problem of computing on which side of the hypersurface the unknown vector lies



Linearly separable vs non-separable classes

- Linearly separable
  - very special and very much desired case of feature space where the classes are separable by a hyperplane
  - all we need to do for perfect classification is to find the right parameters that will separate the classes
- Linearly non-separable classes
  - can be very problematic
  - is dependent of the properties of the feature space
  - we can find a non-linear function that will separate the classes (hard task)

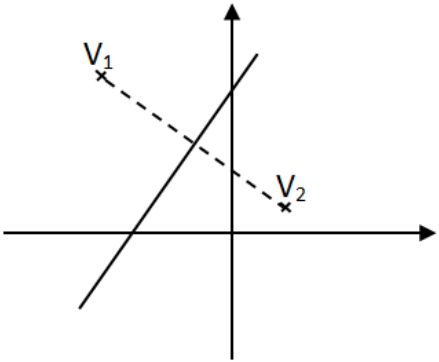
EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
DEPARTMENT OF CYBERNETICS

Loss Functions

- We need to find parameters of optimal hypersurface
- Hinge Loss: $L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta)$
- Cross-entropy (Softmax) Loss:
  - Softmax activation function: $f(\xi)_j = \frac{e^{\xi_j}}{\sum_N e^{\xi_N}}$
  - Cross-entropy loss: $L_{ce} = -\sum_i^N p_i \log \hat{p}_i$
- Triplet Loss: $||x_i^a - x_i^p||_2^2 + \alpha < ||x_i^a - x_i^n||_2^2, \forall (x_i^a, x_i^p, x_i^n) \in T$
- L_2 squared norm (Regression tasks): $L_2 = ||f - y||_2^2$

Minimum distance classifier

- Each class is represented by its representative
- These representatives can be calculated, for example, as an average of all the points from each class (mean vector)
- Advantages: No 'real' training needed, very fast and simple
- Disadvantages: Problems with outliers, Generalization problem



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

k-Nearest Neighbors

- Each image is matched with all images in training data
- Top k with minimum distances are selected
- Majority class of those top k is predicted as output class of the image
- k has to by odd
- Various distance metrics can be used like
  - L1 distance (sum of absolute distance) $E = \sum_{i=1}^{N} |y_i - f(x_i)|$,
  - L2 distance (sum of squares) $E = \sum_{i=1}^{N} (y_i - f(x_i))^2$,
  - etc.
- Voronoi regions for $k = 1$
- Advantages: No 'real' training needed, more robust than minimum distance classifier
- Disadvantages: Necessity to remember the whole training set, huge computational cost

Naive Bayes

- Probabilistic classifier = it makes assumption about data distribution
- Based on Bayes' theorem (Thomas Bayes 1701–1761) with strong (naive) independence assumption between the features

$$P(y_i|\mathbf{X}) = \frac{P(\mathbf{X}|y_i)P(\mathbf{y_i})}{P(\mathbf{X})} \tag{6}$$

- Denominator is constant (can be omitted)
- Bayes' classifier is maximizing probability of correct classification
- maximum a posteriori (MAP) rule $\hat{y} = \arg\max P(y) \prod_{i=1}^{n} P(\mathbf{x}_i|y)$
- Optimizer: Maximum likelihood estimator

Types of Naive Bayes

- Gaussian - assumption of Gaussian distribution of data

$$P(\mathbf{x}_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(\mathbf{x}_i - \mu_y)^2)}{2\sigma_y^2}\right) \tag{7}$$

- Multinomial - the feature vector represent the frequencies of occurrence of certain events
- Bernoulli - Boolean variables

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
DEPARTMENT OF CYBERNETICS

Naive Bayes - properties

- Advantages:
  - If distribution assumption holds true, Naive Bayes can outperform other methods
  - Need only small amount of training data
  - Easy to implement
- Disadvantages:
  - Assumption of independent predictors is almost never true
  - Distribution assumption can be very tricky
  - Covariance shift

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Support Vector Machine

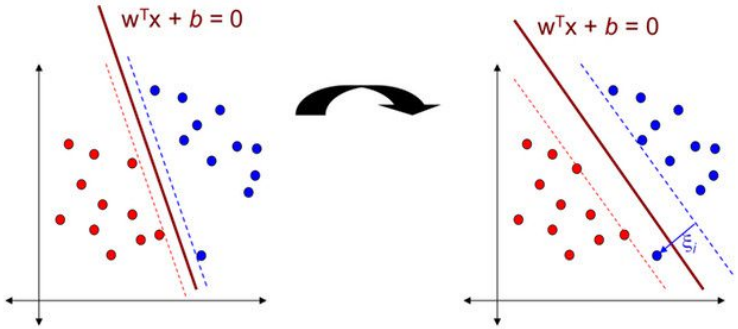
- The best classifier before NN
- Binary linear classifier, more SVMs to perform multi-classs classification, assumes linear separability
- Goal is to design an optimal hyperplane, that correctly classifies all the training vectors
- The criterion: The distance between the boundary and the nearest training vector is maximized
- Score function: $f(\mathbf{x}_i, \mathbf{W}, \mathbf{b}) = \mathbf{W}\mathbf{x}_i + \mathbf{b}$
- Hinge loss

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta) \tag{8}$$

- Optimizer: Lagrangian multiplier method

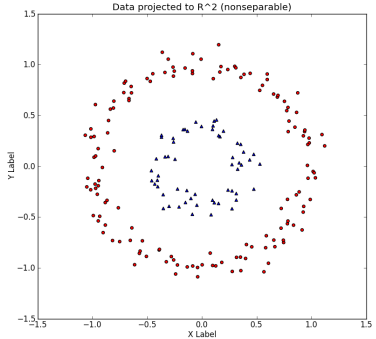
Soft-margin

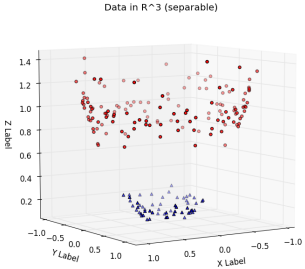
- When classes are not linearly separable
- Goal: Find hyperplane that maximizes the margin and minimizes the number of points within it



Kernel Trick

- Solves problem with linearly non-separable classes
- Transform all patterns (vectors) into a higher dimension
- In this higher dimension the vectors can be linearly separable
- Kernel types: Polynomial, Sigmoid, RBF kernel





SVM - properties

- Advantages:
  - Generally very good results
  - Kernel trick
  - Low risk of over-fitting
- Disadvantages:
  - Choosing a 'good' kernel function is hard
  - Very long training for big datasets

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Softmax Classifier

- Binary linear classifier
- Score function: $f(\mathbf{x}_i, \mathbf{W}, \mathbf{b}) = \mathbf{W}\mathbf{x}_i + \mathbf{b}$
- Cross-entropy loss:

$$L_{ce} = -\sum_{i}^{N} p_i \log \hat{p}_i \tag{9}$$

- Softmax is minimizing the cross-entropy between the estimated class probabilities and the 'true' distribution
- Most popular setup in neural networks classification
- Provides probabilities for each class
- Never truly satisfied

Challenges, codes and examples

- Kaggle
- ImageNet
- Papers with code

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

Thank you for your attention!

Questions?