# Summary

v.v.kitov@yandex.ru

Yandex School of Data Analysis

# Method types

- Label usage:
  - supervised methods use $\{(x_n, y_n)\}_{n=1}^N$
  - unsupervised methods use only$\{x_n\}_{n=1}^N$
- Predicted target (unsupervized):
  - $y \in \mathbb{R}$: feature extraction, dimensionality reduction
    - PCA, non-linear
  - $y \in \{1, 2, ...C\}$: clustering.
    - K-means etc.
- Predicted target (supervized):
  - regression: $y \in \mathbb{R}$
  - classification: $y \in \{1, 2, ...C\}$

# Table of Contents

1. Data preprocessing
2. Supervised prediction methods
3. Extensions
4. Evaluation

# Data visualization

- univariate histograms
- pairwise scatter plots
- in first 2 principal components
- depict $y$ with coordinate, color, marker, point size.

## Common issues

- too much data => subsampling
- outliers => filter them
- missing data => impute it
- imbalanced classes => object reweighting
- train/validation/test have different distributions
    - if adjustable: use random/stratified subsampling
    - if fixed: use adaptation techiques (transfer learning)

# Table of Contents

1. Data preprocessing
2. Supervised prediction methods
3. Extensions
4. Evaluation

# Metric methods

- Assumption: close objects have similar targets
  - “flat costs as average price of similar flats”
- Prediction: by targets of similar objects
- Representatives:
  - KNN, nearest centroids, Parzen window, Nadaraya-Watson regression

# Linear methods

- Assumption: factors linearly add together to form target
  - “#[rooms], #[square meters], distance to city center, etc. sum toghether with coefs to form the price”
- Prediction: by linear functions
- Representatives:
  - regression: linear, ridge, LASSO, robust, SVM regression.
  - classification: SVM, logistic, Perceptron.

# Rule-based methods

- Assumption: target is formed under logical combination of simple rules
  - “if district=central and #[rooms]=1 then price=X, else if ...”
- Prediction: by sequential application of rules
- Representatives: decision trees

# Table of Contents

1. Data preprocessing
2. Supervised prediction methods
3. Extensions
4. Evaluation

# Generalizations

- Feature engineering:
  - centering, scaling
  - $f(x^i)$, $x^i x^j$, $\rho(x - x_{ref})$, etc.
- Metric methods: by different distance functions
- Scalar product dependent: by different kernels $\langle x, z \rangle \to K(x, z)$
  - need to establish Mercer condition for $K(\cdot, \cdot)$

# Ensembles

- Outputs of base models = inputs to other models
- Combination strategies
  - voting
  - averaging
  - step-wise fitting
    - boosting
  - stacking
    - use different datasets for base and aggregation models.
  - iterative stacking with simultaneous fitting
    - neural networks

# Table of Contents

1. Data preprocessing
2. Supervised prediction methods
3. Extensions
4. Evaluation

# General concepts

- Different loss functions give different results
- Loss functions should be derived from business task
    - may be non-symmetric losses

# Classifier evaluation

- Performance on single object:
  - margin: $g_y(x) - \max_{c \neq y} g_c(x)$
  - probability of correct prediction: $p(f(x) = y|x)$
- On the validation set:
  - accuracy
  - ordering of preferences
    - ROC curve, AUC
  - probabilities
    - likelihood function