Statistical Machine Learning - Project

Erik Björk, Jakob Hanson,
Olov Rahm, Simona St

VT 2025

Contents

1 Plan

1.1 From Intro

(i) Explotre and preprocess data

(ii) try some or all classification methods, which are these?

- Logistic Regression
- Discriminant analysis: LDA, QDA
- K-nearest neighbor
- Tree-based methods: classification trees, random forests, bagging
- Boositing

(iii) Which of these are to be "put in producion"?

1.2 From Data analysis task

- Can any trend be seen comparing different hours, weeks, months?
- Is there any diffrence between weekdays and holidays?
- Is there any trend depending on the weather?

1.3 From Implementation of methods

Each group member should implement one family each, who did what shall be clear!

DNNs are encouraged to be implemented, do this if there is time. (DNN is not a thing a group member can claim as their family.)

Implement a naive version, let's do: *Always low_bike_demand*

1.3.1 What to do with each method

1. Implement the method (each person individually)
2. Tune hyper-parameters, discuss how this is done (each person individually)
3. Evaluate with for example cross-validation. Don't use E_{k-fold} (what is that?) (need to do together)
4. (optional) Think about input features, are all relevant? (together)

Before training, unify pre-processing FOR ALL METHODS and choose ONE OR MULTIPLE metrics to evaluate the model. (is it neccesary to have the same for all?, is it beneficial?) Examples:

- accuracy
- f1-score
- recall
- precision

Use same test-train split for ALL MODELS