**Tutorial Nr. 10 – Evaluation & Cross-Validation**

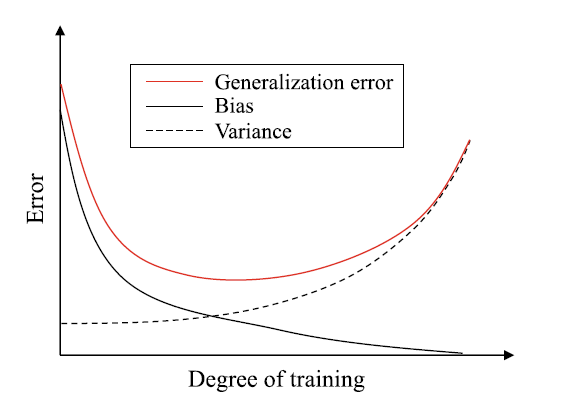
**Timing**

* (particify (5min)) partici.fi/50227978
* Exam
* questions & revision (15min)
* exercise (40min) 🡪 announce which group will present results
* discussion (30min)

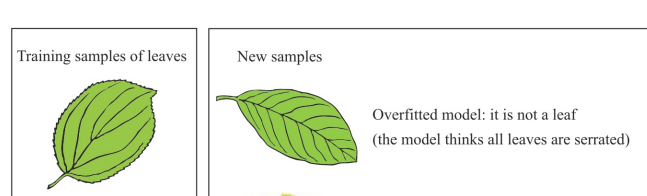
**Questions?**

**Revision**

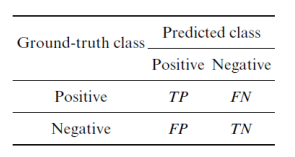
* **test vs training set**
  + **training set:** data on which we “train” a model, given the data it learns about the underlying rules of the data
    - just a proportion of the sample space/population, but should reflect the characteristics of the whole sample space to work well on new samples
  + **test set:** where we test how well the model performs in predicting y on unseen data, given what is has learned with the training data
* **in vs out of sample RSME** 
  + in-sample: how well does the model perform on the training set, e.g. in linear regression
  + out-of-sample: how well does the model generalize to new unseen data, i.e. data outside the training data 🡪 what we’re interested in in ML
    - calculated on test data
  + on test-set; how well your trained model performs on new data
* **bias-variance trade-off**
  + **bias:**
  + inability of the model to capture the true relationship due to modeling assumptions
    - more flexible/complex models have less bias, linear regression more bias
  + **variance:**
    - change in learning performance caused by changes in the training data
    - how model predictions would change if using diff training data
  + **trade-off:** 
    - the more complex & more trained the model, the less bias bc model fits data better; however, variance will increase bc model learns the peculiarities of the data, therefore will vary more between different data sets
    - have to find the optimal trade-off



* **overfitting**
  + when model has learned “too well” on training data 🡪 learns all the peculiarities & generalizes, takes them as general properties that apply to all samples 🡪 poor generalization performance on test set

****

* **true positive /false positive** 
  + true positive: model correctly predicts as positive (presence of characteristic) (TP)
  + false positive: model falsely predicts as positive (FP)
  + true negative: model correctly predicts as negative (absence of characteristic) (TN)
  + false negative: model falsely predicts as negative (FN)
* **accuracy**: proportion of correctly predicted samples (TP+TN/n)
* **precision** (specificity): TP/(TP+FP) 🡪 from all samples being predicted pos, how many are correct
* **recall** (sensitivity): TP/(TP+FN) 🡪 out of all “real” positives, how many have been detected

****

* **resampling method? CV**
* **why do we do CV?**
* **cross-validation**
  + **resampling method:** refit model of interest to samples build from training data to obtain additional info on fitted model
  + used to reduce overfitting by varying samples in test & training set
  + we get multiple evaluation results –> can check whether they are consistent; similar results indicate that we can be more certain about the model's accuracy
  + **process:** 
    - split the data into k-1 training samples, 1 test; train model, get results, evaluate results
      * repeat for k times 🡪 get k results 🡪 average

🡪 more reliable results (less bias) regarding accuracy bc tried on different data

* CV can also be used on training set only, e.g. for model selection (validation)

**Questions?**

**Solution exercise**

1. **Logistic Regression**

* Results cannot interpreted directly because shows the log of the odds of Y being 1 rather than being 0
  + Can only interpret direction & significance
  + Because of non-linearity marginal effects always dependent on values of x variable
* Advantage:
  + Often fits the data better if binary response variable
  + Estimates of response variable between 0 and 1
* AME
  + Reviews: one more review associated on avg w/ decrease in prob of being high rating by 0.24 pp
  + Price: one more $, increase prob by 0.02 pp
  + Gym: having gym, increase prob by 5.8pp

🡪 same results as LPM; Log reg not really better

1. Naïve Bayes

* Column 1 [, 1] ; row 1 (0)🡪 conditional prob of var =1 given that class = 0