Predictive Modelling

Empirical Evaluation

Rita P. Ribeiro

Data Mining I - 2023/2024

Summary

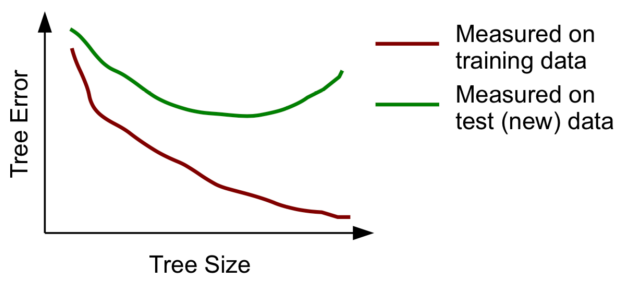
- Hyperparameter Tuning
- Evaluation Methodologies
  - Performance Estimation
  - Experimental Methdologies
- Comparison of Models
  - Statistical Significance Tests
  - Bayesian Approaches

Hyperparameter Tuning

Hyperparameter Tuning

Remember from CART: When to stop growing trees?

- Overall scores keep improving as we grow the tree.
- Still, as we go down in the tree the split decisions are made based on smaller and smaller sets.
- Thus, potentially less reliable decisions are made.



- It is necessary to find the “optimal” tree size, to avoid overfitting.

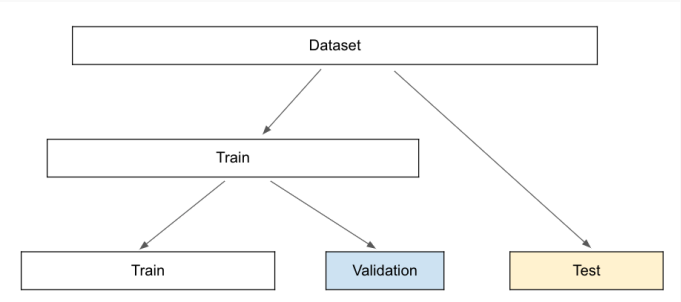
Hyperparameter Tuning

- A hyperparameter is a parameter whose value controls the learning process
- Examples include:
  - k-NN: nr. of neighbors
  - Naive Bayes: Laplace correction
  - Regularized Linear Regression: penalty
  - CART: max. depth, min leaf size, etc.
- Which are the best hyperparameters values for the learning task?
- Conduct a search to tune the hyperparameters
- An experimental methodology is necessary to avoid overfitting

Hyperparameter Tuning

Process:

1) Split the training data: train + validation
   - holdout, cross-validation
2) Based on the model's performance, find the best hyperparameters



3) Train a model with optimal hyperparameters on all training data
4) Test the model on the holdout set

Hyperparameter Tuning

Grid Search

- define grid
  - minimum leaf size: {5,25,50,100}, maximum depth: {3, 5, 10}
- learn and evaluate models for all possible combinations
- choose the best

Random Search

- define domain
  - minimum leaf size: {5,. . . 100}, maximum depth: {3, . . . , 10}
- generate combinations randomly
- learn and evaluate models for the combinations
- choose the best

Hyperparameter Tuning

Other techniques exist

- Bayesian optimization
- Gradient-based optimization
- Evolutionary optimization

Evaluation Methodologies

Performance Estimation

Predictive task

- Learn an approximation to unknown function $Y = f(\mathrm{x})$
- A (training) data set $\{< \mathbf{x_i}, y_i >\}_{i=1}^{N}$, with known values of this mapping
- Performance evaluation criterion - metric of predictive performance (e.g. error rate or mean squared error)
- How to obtain a reliable estimates of the predictive performance of possible solutions to solve the task using the available data set?

Performance Estimation: Reliability of Estimates

- One possible way to obtain an estimate of the performance of a model is to evaluate it on the training data set
- This leads to what is known as a resubstitution estimate of the prediction error
- These estimates are unreliable and should not be used as they tend to be over-optimistic!

Performance Estimation: Reliability of Estimates

Why are the resubstitution estimates unreliable?

- Models are obtained with the goal of optimizing the selected prediction error statistic on the given data set
- In this context, it is expected that they get good scores!
- The given data set is just a sample of the unknown distribution of the problem being tackled
- What we would like is to have the performance of the model on this distribution
- As this is usually impossible the best we can do is to evaluate the model on new samples of this distribution

Performance Estimation: Main Goal

- Obtain a reliable estimate of the expected prediction error of a model on the unknown data distribution
- In order to be reliable it should be based on evaluation on unseen cases - a test set

The golden rule

- *The data used for evaluating (or comparing) any models cannot be seen during model development.*

Performance Estimation: Experimental Methodology

- Ideally, we should repeat the testing several times
- Collect a series of scores and provide as estimate the average of these scores, together with its standard error
- In summary:
  - calculate the sample mean prediction error on the repetitions as an estimate of the true population mean prediction error
  - complement this sample mean with the standard error of this estimate

Performance Estimation: Experimental Methodology

An experimental methodology should:

- Allow obtaining several prediction error estimates of a model, $e = \{e_1, e_2, \cdots, e_k\}$
- Such that we can calculate a sample mean prediction error

$$\overline{e} = \frac{1}{K} \sum_{i=1}^{k} e_i$$

- And also the respective standard error of this estimate, based on the sample standard deviation of e, S_E

$$SE(\overline{e}) = \frac{S_e}{\sqrt{k}} = \frac{\sqrt{\frac{1}{k-1} \sum_{i=1}^{k} (e_i - \overline{e})^2}}{\sqrt{k}}$$

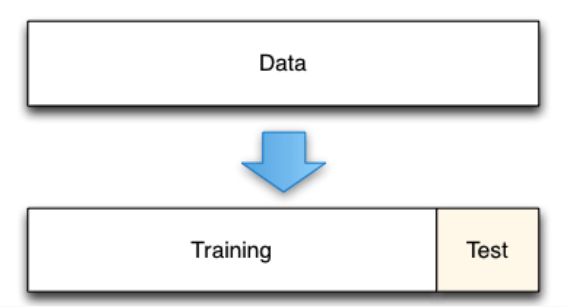
Experimental Methodologies

- Holdout Method
- Random Subsampling
- k-fold Cross Validation
- Leave One Out Cross Validation
- Bootstrap Method

Experimental Methodologies

Holdout Method

- It consists on randomly dividing the available data sample in two sub-sets
  - one used for training the model;
  - and the other for testing/evaluating it
  - a frequently used proportion is 70% for training and 30% for testing



Experimental Methodologies

Holdout Method (cont.)

- Preferred for very large data samples
- Small data sample
  - danger of either having a too small test set (unreliable estimates as a consequence)
  - or removing too much data from the training set (worse model than what could be obtained with the available data)
  - if only one prediction error score is obtained - no average score nor standard error

Experimental Methodologies

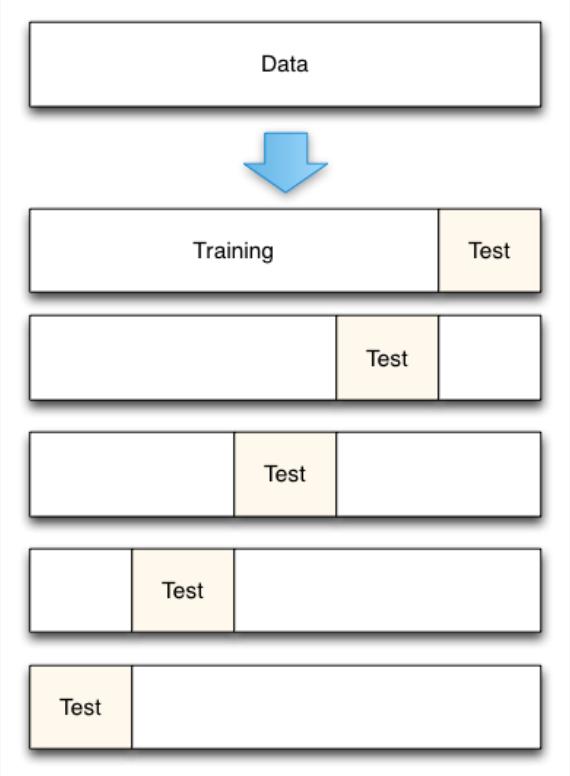
Random Subsampling

- Variation of holdout method and it simply consists of repeating the holdout process several times by randomly selecting the train and test partitions
- Has the same problems as the holdout with the exception that we already get several scores and thus can calculate means and standard errors
- If the available data sample is too large the repetitions may be too demanding in computation terms

Experimental Methodologies

k-fold Cross Validation

- The idea of k-fold Cross Validation (CV) is similar to random subsampling
- It essentially consists of k repetitions of training on part of the data and then test on the remaining
- The diference lies on the way the partitions are obtained
- $\hat{e}_{CV} = \bar{e} \pm SE(\bar{e})$



Experimental Methodologies

Advantage of Cross Validation

- We have the guarantee that each example will be used at least once for training and other for testing.

Stratified k-fold Cross Validation

- If it is expected that the learning algorithm to sensitive to the target variable distribution
- Each fold has roughly the same distribution

Leave One Out Cross Validation (LOOCV)

- Similar idea to k-fold Cross Validation (CV) but in this case on each iteration a single case is left out of the training set
- Equivalent to n-fold CV, where n is the size of the data set.

Experimental Methodologies

Bootstrap Method

- Train a model on a random sample of size n with replacement from the original data set (of size n)
  - Sampling with replacement means that after a case is randomly drawn from the data set, it is "put back on the sampling bag"
  - Several cases will appear more than once on the training data
  - On average only 63.2% of all cases will be on the training set
- Test the model on the cases that were not used on the training set
- Repeat this process many times (typically around 200)

Bootstrap Method (cont.)

- The average of the scores on these repetitions is known as the bootstrap estimate, but it is an optimistic estimate because of the overlap between training and test cases.
- Alternative: .632 bootstrap estimate, obtained by

$$\hat{e}_{.632} = .632 \times e_0 + .368 \times e_r$$

where

  - e_0 is the leave-one-out bootstrap estimate, i.e. the average of the error for each case i is based on the bootstrap samples in which i does not appear (pessimistic estimate)
  - e_r is the resubstitution estimate (optimistic estimate)
- This method is more appropriate when the data set is small

Comparison of Models

Comparison of Models

- In order for two models to be comparable they have to refer to same training and test datasets.
- Any of the previous performance estimation methods can be used.
- But the same has to be used for all the models in comparison.
- Typically, the goal is confirm whether a given algorithm performs better w.r.t a baseline algorithm.

Comparison of Models

Example

- Accuracy estimates obtained by a 10-fold CV of 3 models

Fold	**M1**	**M2**	**M3**
1	0.735	0.875	0.976
2	0.627	0.719	0.663
3	0.732	0.897	0.988
4	0.573	0.736	0.845
5	0.721	0.844	0.848
6	0.688	0.607	0.924
7	0.570	0.609	0.691
8	0.689	0.927	0.663
9	0.686	0.734	0.696
10	0.610	0.887	0.855
avg	0.663	0.783	0.815
stdev	0.063	0.119	0.128

- We can say that **M3** is better.
- But, are the models' performance really different?

Comparison of Models

- For each pair of models and for each fold, we calculate the difference between the accuracy values.

Fold	**M1-M2**	**M1-M3**	**M2-M3**
1	-0.140	-0.241	-0.100
2	-0.092	-0.036	0.056
3	-0.165	-0.256	-0.092
4	-0.164	-0.272	-0.109
5	-0.123	-0.128	-0.005
6	0.081	-0.236	-0.317
7	-0.039	-0.121	-0.082
8	-0.238	0.0256	0.264
9	-0.048	-0.010	0.038
10	-0.277	-0.245	0.032
avg	-0.120	-0.152	-0.032
stdev	0.103	0.113	0.150

- Are the observed differences statistically significant?

Statistical Significance

Null Hypothesis Statistical Tests (NHST)

- Test if some result is unlikely to have occurred by chance
- The null hypothesis (H_0): there is no difference among a set of models, i.e. the true difference is 0 and any differences in performance are attributed to chance
- This hypothesis is rejected if the result of the significance test has a *p*-value less than a certain selected threshold α for significance level.
- *p*-value is the probability of observing a difference as large as the sample difference given H_0.
  - If $p - value < \alpha$, then H_0 is rejected with $(1 - \alpha)$ confidence.
  - Typically signficance level is 0.01 or 0.05, i.e. 99% or 95% confidence, respectively.

Statistical Significance

- Different statistical significance tests can be used.
- Compare different models in a specific data set (prediction task)
  - Paired *t*-test
  - Wilcoxon's signed-rank test
- Compare different models across multiple data sets (prediction tasks)
  - Friedman test and post-hoc tests
    - Nemenyi test,
    - Bonferroni-Dunn test

Paired Comparisons on a Single Task

Paired *t*-test

- Parametric test that can be used to compare two paired samples
- Assumptions:
  - data are paired and come from the same population
  - data collected from a representative, randomly selected portion of the total population
  - sample is drawn from a population with a Normal (Gaussian) distribution
  - the size of the sample is reasonably large.

Paired Comparisons on a Single Task

Paired *t*-test (cont.)

- Test procedure for $\{m_{1,i}, m_{2,i}\}_{i=1}^{N}$:
  - Find the difference between pairs $\{d_i\}_{i=1}^{N}$
  - The assumption is that the difference between two normally distributed varaiables is also normally distributed
  - H_0 is that the differences have mean 0 and unknown standard deviation.
  - Calculate the *p*-value
  - If $p - value < \alpha$, then H_0 is rejected, which means that the performance of the models is statistically different.

Paired Comparisons on a Single Task

Paired *t*-test (cont.)

Fold	**M1-M2**	**M1-M3**	**M2-M3**
1	-0.140	-0.241	-0.100
2	-0.092	-0.036	0.056
3	-0.165	-0.256	-0.092
4	-0.164	-0.272	-0.109
5	-0.123	-0.128	-0.005
6	0.081	-0.236	-0.317
7	-0.039	-0.121	-0.082
8	-0.238	0.0256	0.264
9	-0.048	-0.010	0.038
10	-0.277	-0.245	0.032
avg	-0.120	-0.152	-0.032
stdev	0.103	0.113	0.150
p-value	0.0049	0.0022	0.5237

- The differences **M1**-**M2** and **M1**-**M3** are found significant at the $\alpha = 0.01$ level, i.e. 99% confidence level (and thus, also at $\alpha = 0.05$ level, i.e. 95% confidence level).

Paired Comparisons on a Single Task

Wilcoxon's signed-rank test

- Non-parametric test based on ranking information
  - differences are still taking into account but only qualitatively, absolute magnitudes are ignored
  - as it does not assume normal distribution, the outliers (exceptionally good/bad performances) have less effect on the Wilcoxon than on the t-test.

Paired Comparisons on a Single Task

Wilcoxon's signed-rank test (cont.)

- Test procedure for $\{m_{1,i}, m_{2,i}\}_{i=1}^{N}$:
  - Find the difference between pairs $\{d_i\}_{i=1}^{N}$
  - Record the sign of the difference and the absolute value of the difference
  - Rank the absolute differences from the smallest to the largest.
  - Re-attach the signs of differences to the respective ranks to obtain signed ranks, then obtain
    - $R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$
    - $R^+ = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$
  - Take the smallest of these sums and find the p-value that rejects H_0 at a significance level (α)

Paired Comparisons on a Single Task

Wilcoxon's signed-rank test (cont.)

Fold	**M1-M2** (rank)	**M1-M3** (rank)	**M2-M3** (rank)
1	-0.140 (6)	-0.241 (7)	-0.100 (7)
2	-0.092 (4)	-0.036 (3)	0.056 (4)
3	-0.165 (8)	-0.256 (9)	-0.092 (6)
4	-0.164 (7)	-0.272 (10)	-0.109 (8)
5	-0.123 (5)	-0.128 (5)	-0.005 (1)
6	0.081 (3)	-0.236 (6)	-0.317 (10)
7	-0.039 (1)	-0.121 (4)	-0.082 (5)
8	-0.238 (9)	0.0256 (2)	0.264 (9)
9	-0.048 (2)	-0.010 (1)	0.038 (3)
10	-0.277 (10)	-0.245 (8)	0.032 (2)
R^+	3	2	18
R^-	52	53	37
p-value	0.0098	0.0059	0.375

- The differences **M1**-**M2** and **M1**-**M3** are found significant at the $\alpha = 0.01$ level, i.e. 99% confidence level.

Paired Comparisons on a Single Task

Parametric or non parametric-test?

- Large samples
  - The central limit theorem ensures that parametric tests work well with large samples even if the population is non-Gaussian.
- Small samples
  - Using a parametric test with data from non-Gaussian populations, you can't rely on the central limit theorem, so the *p*-value may be inaccurate.
  - Using a non-parametric test is, probably, a safer option.
  - However, a non-parametric test has less power to detect a real effect than the parametric test if all the assumptions underlying the parametric test are satisfied.

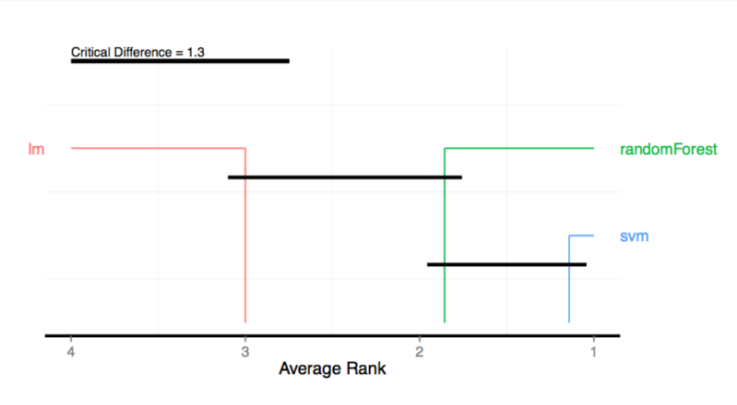
Multiple Comparisons on Multiple Tasks

Demsar (2006) recommended procedure

- Step 1: Friedman test
  - H_0: all models are equivalent and so their rankings across the tasks are equal
- Step 2: If H_0 is rejected then we can move to
  - Nemenyi post-hoc test: paired comparisons among all pairs of models
    - H_0: there is no significant difference among the ranks of a certain pair of models
  - Bonferroni-Dunn post-hoc test: paired comparisons against a baseline
    - H_0: there is no significant difference among the ranks of a certain model and the baseline

Multiple Comparisons on Multiple Tasks

- Critical Difference (CD) Diagrams
  - shows the average rank of each model
  - average rank differences that are not statistically significant are connected
  - it can be obtained for Nemenyi or Bonferroni-Dunn post-hoc tests



Bayesian Approaches

How probable it is that model M1 is better than M2 by more than 1%?

- Bayesian sign test compares two models on multiple data sets by determining the posterior probability distribution of the performance differences between two models (Benavoli et al. (2015))
- Region of Practical Equivalence (ROPE): a region in the probability density space where the two classifiers are practically equivalent.
- Probability of M1 being practically better/equivalent/worse than M2.

References

References

Aggarwal, Charu C. 2015. *Data Mining, the Texbook*. Springer.

Benavoli, Alessio, Giorgio Corani, Francesca Mangili, and Marco Zaffalon. 2015. “A Bayesian Nonparametric Procedure for Comparing Algorithms.” In *Proceedings of the 32nd International Conference on Machine Learning*, edited by Francis Bach and David Blei, 37:1264–72. Proceedings of Machine Learning Research. Lille, France: PMLR. https://proceedings.mlr.press/v37/benavoli15.html.

Demsar, Janez. 2006. “Statistical Comparisons of Classifiers over Multiple Data Sets.” *Journal of Machine Learning Research* 7: 1–30.

Flach, Peter. 2012. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511973000.

Gama, João, André Carlos Ponce de Leon Ferreira de Carvalho, Katti Faceli, Ana Carolina Lorena, and Márcia Oliveira. 2015. *Extração de Conhecimento de Dados: Data Mining -3rd Edition*. Edições Sílabo.

Torgo, Luís. 2017. “Data Mining i Course.” Slides.