COMP24112: Machine Learning

Chapter 3: Machine Learning Experiments III

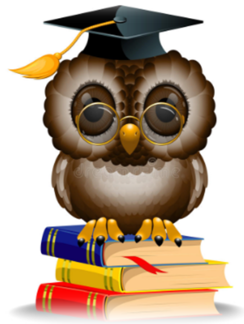
Dr. Tingting Mu

Email: tingting.mu@manchester.ac.uk

Content

MANCHESTER
1824
The University of Manchester

- Hypothesis and model.
- Evaluate hypothesis:
    - Sample error close to the true error?
- Compare hypotheses:
    - Sample error difference close to true error difference?



Hypothesis

- **Remember?** A hypothesis refers to a prediction made by a trained machine learning model.
- Often we talk about the context where a hypothesis refers to a model trained on a sample set:

$$Hypothesis\ A\left(T\right): \text{the model A trained on sample set T.}$$

- For example:
  - Hypothesis 1: A 5-NN classifier trained on sample set A.
  - Hypothesis 2: A 3-NN classifier trained on sample set B.
  - Hypothesis 3: A 6-NN regressor trained on sample set C.

Model Evaluation

- True error of a model *A*: Expectation of the true error of the hypothesis *A(T)* with randomly drawn training set *T*.

$$E_{T \subset D}\left\{ \underline{error_D\left(A\left(T\right)\right)} \right\}$$

Hypothesis evaluation:
- Train your model using training data T.
- Estimate the true error using a new test data E.

Model Evaluation

- True error of a model *A*: Expectation of the true error of the hypothesis *A(T)* with randomly drawn training set *T*.

$$\underline{E_{T\subset D}}\left\{error_D\left(A\left(T\right)\right)\right\}$$

Model evaluation:
- Approximate the expectation, by running multiple training-testing trials and average the test error rates.

- This motivates evaluation methods like random subsampling, k-fold CV, LOO, bootstrap, based on multiple trials of training and testing.

Model Comparison

- Often, we are interested in evaluating and comparing machine learning models (approaches, algorithms).
  - For example, to compare 5-NN,10-NN and 1-NN for classification.

- Performance difference between models:

$$E_{T\subset D}\left\{\underline{error_D\left(A\left(T\right)\right)-error_D\left(B\left(T\right)\right)}\right\}$$

Compare two hypotheses first.

Hypothesis Evaluation

- Given a trained model, we usually use a new set of samples to estimate its performance.

- **Question**: *How good an estimate of the true error is provided by the sample error?*

- Check this using **confidence interval**!

Confidence Interval for Classification

- You have computed the sample classification error, using a set of *n* samples.
- Confidence interval tells you

With p probability, the true error lies in the interval of

$$error_D \in \left[error_S - \underline{a}, error_S + \underline{a}\right].$$

- We wish to to have a small *a* for a more precise estimate, and a large *p* for higher confidence.
- How do you compute *a* given the chose *p*?

Confidence Interval for Classification

- You can compute the value of *a* using the equation and table below:

$$a = z_p \sqrt{\frac{error_s\left(1 - error_s\right)}{n}}$$

Table of z_p value for two-sided p confidence interval.

Confidence level: p	50%	68%	80%	90%	95%	98%	99%
Constant: z_p	0.67	1.00	1.28	1.64	1.96	2.33	2.58

With p probability, the true error lies in the interval of $error_D \in \left[error_S - a, error_S + a\right]$.

Summary

With p probability, the true error lies in the interval of

$$error_D \in \left[error_s - z_p \sqrt{\frac{error_s(1-error_s)}{n}}, error_s + z_p \sqrt{\frac{error_s(1-error_s)}{n}} \right].$$

Confidence level p%	50%	68%	80%	90%	95%	98%	99%
Constant z_p	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Comments:

- This confidence interval is an approximate.
- It works pretty well for over 30 samples and with sample error not too close to 0 or 1.

Compare Two Hypotheses

- **Classifier A**: error rate computed using a set of n_1 samples, denoted by $error_{s1}(A)$.
- **Classifier B**: error rate computed using a set of n_2 samples, denoted by $error_{s2}(B)$.
- **Fact**: $error_{s1}(A)-error_{s2}(B)>0$

Note: sample errors of the two classifiers can be based on different sample set.

- Question:

Given that classifier A has higher sample error than classifier B

$$error_{s1}(A)-error_{s2}(B)>0,$$

what is the probability C that classifier A has higher true error than classifier B?

$$error_D(A)-error_D(B)>0$$

z-Test

Given that classifier A has higher sample error than classifier B: $error_{s1}(A)-error_{s2}(B)>0$, what is the probability C that classifier A has higher true error than classifier B: $error_D(A)-error_D(B)>0$?

- You can use **z-Test** for this.
  - Step 1: Compute a quantity z_p as below:

$$z_p = \frac{d}{\sigma}, \text{ where}$$

$$d = \left| error_{s1}(A) - error_{s2}(B) \right|$$

$$\sigma = \sqrt{\frac{error_{s1}(A)\left[1 - error_{s1}(A)\right]}{n_1} + \frac{error_{s2}(B)\left[1 - error_{s2}(B)\right]}{n_2}}$$

z-Test

– Step 2: Look up the table below to get the confidence value *p.*

Table of z_p value for two-sided p confidence interval.

Confidence level: p	50%	68%	80%	90%	95%	98%	99%
Constant: z_p	0.67	1.00	1.28	1.64	1.96	2.33	2.58

– Step 3: Compute the final probability by

$$C = 1 - \frac{(1-p)}{2}$$

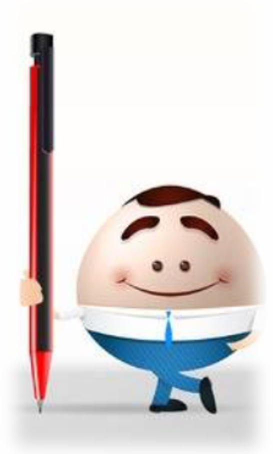
Comments on z-Test

- It only compares two hypotheses at a time.
- Hypotheses can be tested on different sets of samples.
- The approximation works well for test sets containing over 30 samples.

Chapter 3 Summary: A, B and C

MANCHESTER 1824
The University of Manchester

- Measure classification and regression performance using samples.
  - Classification: Accuracy/error, confusion matrix, precision, recall, F1 score, specificity
  - Regression: RMSE, MAE, MAPE, R^2 score
- Issues of evaluation with limited data
  - Sample error and true error
  - Bias issue and variance issue
  - Never train, test and select model using the same sample set.



- Machine learning experiments
  - Data split strategies: holdout, random subsampling, k-fold CV, LOO, bootstrap
  - Model training, evaluation and selection
- Bias and variance decomposition
- Evaluate and compare hypotheses
  - Confidence interval
  - Z-score test