Tampere University

Confidence intervals
Why report them and how to calculate them?

Annamaria Mesaros

Tampere University

Partially based on sections 1 and 2 in Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning by Sebastian Raschka

Model performance

“Outperforming current state of the art” - How measured?

Performance evaluation and comparison is very important

Performance comparison is always on limited data

Confidence intervals give an idea of the uncertainty of the reported performance

What is a Confidence Interval?

A confidence interval is a method that computes an upper and a lower bound around an estimated value

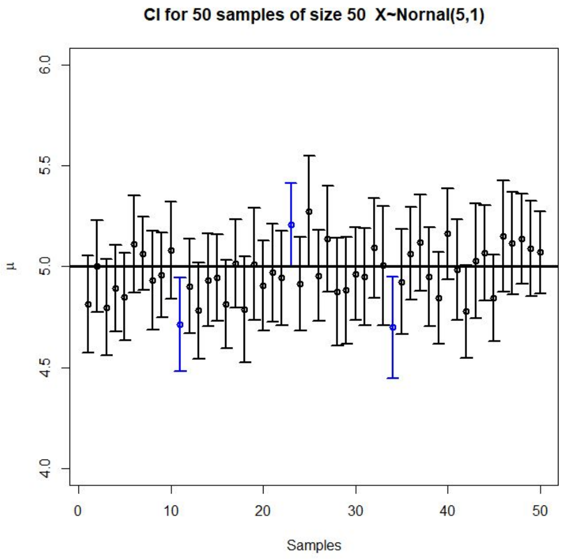
e.g. sample mean

- calculated from a sample (finite!) drawn from an unknown population
- estimated as the mean of the sample, to characterize the entire population
- but is not exactly the same! If we draw a different sample, we may obtain a different estimate

What is a Confidence Interval?

95% confidence interval

- assume we have access to the population (not happening in real life) and we know the exact mean
- draw samples from the population, estimate the mean of these samples and their 95% CIs
- 95% of the calculated CIs will contain the true value



"There is a 95% probability that the 95% confidence interval calculated from a given future sample will cover the true value of the population parameter."

Confidence interval in machine learning

We calculate a model's performance on a test dataset

We interpret it as an estimated generalized accuracy

Expect a similar performance on different samples of a very large test dataset same distribution

The 95% CI gives us some uncertainty on how accurate this estimate is

Confidence interval in machine learning

We calculate a model's performance on a test dataset

We interpret it as an estimated generalized accuracy

Expect a similar performance on different samples of a very large test dataset same distribution

The 95% CI gives us some uncertainty on how accurate this estimate is

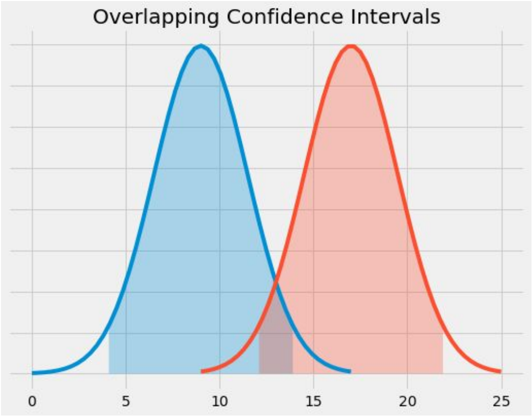
A 95% confidence interval does not mean that there is a 95% probability that the true value is within the interval

Statistical significance

“Outperforming current state of the art”

We can say that the difference of two measurements is statistically significant if **confidence intervals do not overlap**.

!!! We *cannot* say that results are ***not*** **statistically significant** if confidence intervals overlap. (hypothesis testing, sample size, etc)



By Eugine Kang
https://medium.com/@kangeugine/overlapping-and-difference-confidence-intervals-d163a86b3a00

Normal approximation

Confidence interval for an estimated parameter (let’s say the the sample mean) assuming a normal distribution:

$$\bar{x} \pm z \times \mathrm{SE}$$

where

- z is the z value (the number of standard deviations that a value lies from the mean of a standard normal distribution; usually looked up in tables);
- SE is the standard error of the estimated parameter (here: sample mean)

$$\mathrm{SE} = \sqrt{\frac{1}{n}\mathrm{ACC}_{\mathrm{test}}\left(1 - \mathrm{ACC}_{\mathrm{test}}\right)}$$

Accuracy = a proportion of success (Binomial proportion success interval)

So the confidence interval is $\mathrm{ACC}_{\mathrm{test}} \pm z\sqrt{\frac{1}{n}\mathrm{ACC}_{\mathrm{test}}\left(1 - \mathrm{ACC}_{\mathrm{test}}\right)}$

Bootstrapping and empirical CIs

Useful when we don't have access to the sample's distribution (the behavior of our measure)

Bootstrap = generate new data from a population by repeated **sampling** from the original dataset **with replacement**

Holdout (folds) = sampling without replacement.

- Estimate of the model's prediction accuracy

Bootstrapping and empirical CIs

Given a dataset of size *n*:

- For *b* bootstrap rounds:
  - Draw one single instance from this dataset and assign it to the *j*th bootstrap sample.
  - Repeat this step until the bootstrap sample has size *n* (the size of the original dataset)
  - Certain examples may appear more than once in a bootstrap sample and some not at all.
- Fit a model to each of the *b* bootstrap samples and compute the accuracy.
- Compute the model accuracy as the average over the *b* accuracy estimates

Figure from: Sebastian Raschka, Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning

A "good" number of bootstrap samples is considered to be 200

VERY EXPENSIVE!!

Bootstrapping and empirical CIs

- Take multiple samples from a single random sample and estimate the sampling distribution

$$\mathrm{ACC}_{boot} = \frac{1}{b}\sum_{i=1}^{b}\mathrm{ACC}_i$$

- If they follow a normal distribution, use the same formula for SE

$$\mathrm{SE}_{boot} = \sqrt{\frac{1}{b-1}\sum_{i=1}^{b}(\mathrm{ACC}_i - \mathrm{ACC}_{boot})^2}$$

- Then calculate confidence interval:

$$\mathrm{ACC}_{boot} \pm t \times \mathrm{SE}_{boot}$$

Originally, the bootstrap method aims to determine the statistical properties of an estimator when the underlying distribution was unknown and additional samples are not available

Bootstrapping (2)

If no assumption on distribution: use percentile method [Efron, 1981]

- $ACC_{lower} = \alpha_1$th percentile of the ACC_{boot} distribution
- $ACC_{upper} = \alpha_2$th percentile of the ACC_{boot} distribution

Where $\alpha_1 = \alpha$ *and* $\alpha_2 = 1 - \alpha$

- α is the degree of confidence for computing the $100 \times (1 - 2 \times \alpha)$ confidence interval.

For a 95% confidence interval, $\alpha = 0.025$, which gives the 2.5th and 97.5th percentiles of the *b* bootstrap samples distribution as the upper and lower confidence bounds.

Jackknife

Resampling the test set!

- Leave-one out: calculate performance of the model on test set by leaving out, in turn, one test item
- Similar to the holdout method in training (leave-one-out cross-validation procedure)
- Based on the obtained sample, estimate standard error and confidence intervals

Advantages:

- The model is fixed, we only need the model output
- No retraining is necessary
- No assumptions on the distribution of the sample (metric)
- Allows direct comparison with published work if the authors have reported CIs (same test set)

Retraining models with different random seed

Common procedure: retrain a model with different random seeds, then compute CI based on them

Assuming normally distributed samples, use formula (t-value instead of z-value because low number of samples)

What does this CI tell us? information on the stability of the model

Retraining models with different random seed

Can be used to compare two models m_1 and m_2

- testing the difference of proportions based on the normal approximation (assuming unequal variances)

$$\left(\overline{ACC}_{\mathrm{m1}} - \overline{ACC}_{\mathrm{m2}}\right) \pm t\sqrt{\frac{\mathrm{SD}^2_{\mathrm{m1}}}{n_{\mathrm{m1}}} + \frac{\mathrm{SD}^2_{\mathrm{m2}}}{n_{\mathrm{m2}}}}$$

- if the calculated 95% CI does not contain 0, the performance of the models is statistically significant at alpha=0.05

Disadvantage: needs retraining both models multiple times

- expensive
- only applicable if you have both models

Retraining models with different random seed

Can be used to compare two models m_1 and m_2

- testing the difference of proportions based on the normal approximation (assuming unequal variances)

$$\left(\overline{ACC}_{\mathrm{m1}} - \overline{ACC}_{\mathrm{m2}}\right) \pm t \sqrt{\frac{\mathrm{SD}^2_{\mathrm{m1}}}{n_{\mathrm{m1}}} + \frac{\mathrm{SD}^2_{\mathrm{m2}}}{n_{\mathrm{m2}}}}$$

- if the calculated 95% CI does not contain 0, the performance of the models is statistically significant at alpha=0.05

Disadvantage: needs retraining both models multiple times

- expensive
- only applicable if you have both models

McNemar Test is a much better choice for comparing two classifiers [McNemar, 1947]

Take away

- Reporting results is more complete if CIs are given with the performance

  You would also like to to know if the 2% or 0.2 [whatever unit] improvement you obtained matters
- When doing classification (accuracy) there is an easy formula, so you have no excuse
- For any other metric, the jackknife procedure is very fast and simple (so you have no excuse)