# Bayesian Hypotheses Testing

Jakub Repický

Faculty of Mathematics and Physics,
Charles University

Institute of Computer Science,
Czech Academy of Sciences

Selected Parts of Data Mining
Jan 19 2018, Prague

Based on tutorial by Benavoli et al.
`http://ipg.idsia.ch/tutorials/2016/bayesian-tests-ml/`

Experiments:

- Comparing Adaboost (ada) vs. Gradient boosting classifier (gbc)
- scikit-learn implementation
- `max_depth=1, n_estimators=100`
- `learning_rate=1.0` (gbc)

# Data

Table: 27 UCI data sets

| | Name | Size | No. of features |
|---|---|---|---|
| 0 | heart-statlog | 270 | 13 |
| 1 | mushroom | 5644 | 22 |
| 2 | segment | 2310 | 19 |
| 3 | cleveland-14-heart-disease | 296 | 13 |
| 4 | zoo | 101 | 17 |
| | ... | | |
| 23 | ionosphere | 351 | 34 |
| 24 | pima_diabetes | 768 | 8 |
| 25 | vote | 232 | 16 |
| 26 | vehicle | 846 | 18 |

# Procedure of NHST

1. State the null and the alternative hypotheses $H_0$ and $H_1$
2. Based on statistical assumption about data, choose a statistical test
3. Under the null hypothesis, the test statistic $T$ follows a known probability distribution
4. Calculate observed test statistic $t(\boldsymbol{x})$
5. Calculate the probability that $T$ is "more extreme" than observed $t(\boldsymbol{x})$ (the $p$-value)
6. If $p < \alpha$, reject $H_0$

# Correlated $t$-test

- Used to test two algorithms on one data set
- Calculates a score (e. g., accuracy) on $p$ runs of $k$-fold cross-validation
- Sample size: $n = pk$
- Observations: $\boldsymbol{x} = (x_i)_{i=1}^{n}$, the score differences on each fold
- The standard $t$-test assumes $x_i$ to be independently, identically and normally distributed
- Correlated $t$-test accounts for correlations between $x_i$, $x_j, i \neq j$ due to cross-validation

# Correlated $t$-test (II)

The test statistic:

$$t(\boldsymbol{x}, \mu) = \frac{\bar{\boldsymbol{x}} - \mu}{\sqrt{\hat{\sigma}^2 \left( \frac{1}{n} - \frac{\rho}{1-\rho} \right)}}$$

- $t$ follows Student's distribution with $n - 1$ degrees of freedom
- $\rho$ – correlation between results from overlapping training sets
- $\frac{\rho}{1-\rho} = \frac{n_{\mathrm{te}}}{n_{\mathrm{tr}}}$ – a heuristic for the correlation correction parameter (Nadeau and Bengio, 2003)
- Two-sided test: $H_0 : \mu = 0,\ H_1 : \mu \neq 0$
- One-sided test: $H_0 : \mu \leqslant 0,\ H_1 : \mu > 0$

# Example

Table: $p$-values of the two-sided correlated $t$-test. 14 out of 27 results are significant at $\alpha = 0.05$.

| | Name | p-val |
|---|---|---|
| 0 | heart-statlog | 0.51 |
| 1 | mushroom | 0.00* |
| 2 | segment | 0.00* |
| 3 | cleveland-14-heart-disease | 0.42 |
| 4 | zoo | 0.00* |
| | ... | |
| 23 | ionosphere | 0.23 |
| 24 | pima_diabetes | 0.29 |
| 25 | vote | 0.39 |
| 26 | vehicle | 0.00* |

# Wilcoxon signed-rank test

- Used to compare two classifiers on multiple data sets
- Counts ranks of differences, not their magnitudes
- $z_i$ – the mean score difference on $i$th data set, $i = 1, \ldots, q$
- $z_i$ assumed to be i.i.d. samples from a symmetric distribution

# Wilcoxon signed-rank test (II)

- The test statistic is

$$t = \min\left\{ \sum_{i:z_i>0} \mathrm{rank}(|z_i|) + \frac{1}{2}\sum_{i:z_i=0} \mathrm{rank}(|z_i|), \right.$$
$$\left. \sum_{i:z_i<0} \mathrm{rank}(|z_i|) + \frac{1}{2}\sum_{i:z_i=0} \mathrm{rank}(|z_i|) \right\}$$

- Critical value tables exist for $q$ small enough, e. g., $q < 25$
- Otherwise $w = \frac{t-\frac{1}{4}q(q+1)}{\sqrt{\frac{1}{24}q(q+1)(2q+1)}}$ follows an approximately normal distribution

# Example

Wilcoxon signed-rank test of mean accuracy difference between ada and gbc:

$$w = 120,\ p\text{-value} = 0.10.$$

## $p$-value not what researchers want

- $p$-value is not the probability of the null hypothesis

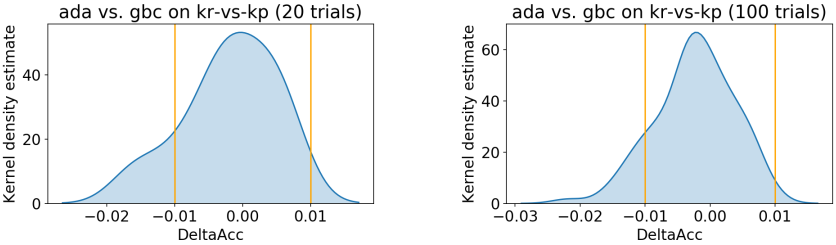
$$p(T > t(\boldsymbol{x})|H_0) \neq p(H_0|\boldsymbol{x})$$

- Similarly, $1 - p$ is not the probability of the alternative hypothesis

$$p(T \leqslant t(\boldsymbol{x})|H_0) \neq p(H_1|\boldsymbol{x})$$

# $p$-value depends on sample size

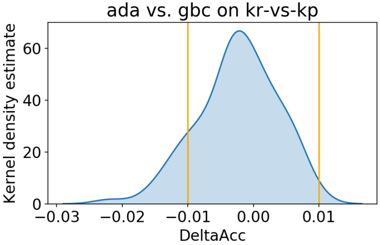
- The difference between classifiers is never zero
- Arbitrarily small effects can be confirmed on large enough samples



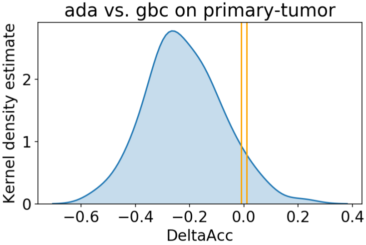

$20$ trials, $p$-value$= 0.24$ $\quad$ $100$ trials, $pr < 10^{-3}$

# NHST cannot measure effect size

- Statistical significance does not imply practical significance




$p < 10^{-3}$




$p < 10^{-3}$

# And more. . .

- If null hypothesis is not rejected, the result is inconclusive
- Significance level cannot be reasonably decided
- NHST assumes certain sampling intentions

# Bayesian analysis

Bayesian inference:

1. Formulating a joint probability model of observable data $\boldsymbol{x}$ and unknown parameters $\theta$:

$$p(\theta, \boldsymbol{x}) = p(\boldsymbol{x}|\theta)p(\theta)$$

2. Infering $\theta|\boldsymbol{x}$ by Bayes' theorem:

$$p(\theta|\boldsymbol{x}) = \frac{p(\theta, \boldsymbol{x})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})}$$

3. Summarizing the posterior distribution

# Bayesian correlated $t$-test

Likelihood:

$$\boldsymbol{x} \mid \mu, \tau \sim \mathrm{MVN}(\mu \mathbf{1}, \Sigma)$$

$$\Sigma = \begin{pmatrix} 1/\tau & \rho/\tau & \cdots & \rho/\tau \\ \rho/\tau & 1/\tau & \cdots & \rho/\tau \\ \vdots & \vdots & \ddots & \vdots \\ \rho/\tau & \rho/\tau & \cdots & 1/\tau \end{pmatrix}$$

Prior:

$$\begin{aligned} \mu, \tau &\sim \mathrm{NormalGamma}(\mu_0, k_0, a, b) \\ \mu \mid \tau &\sim \mathcal{N}(\mu_0, k_0/\tau) \\ \tau &\sim \mathrm{Gamma}(a, b) \end{aligned}$$

# Bayesian correlated $t$-test (II)

- NormalGamma is conjugate to MVN
- The posterior is a NormalGamma distribution
- Marginalizing out precision $\tau$, the posterior of $\mu$ is a Student $t$-distribution
- For $\mu_0 = 0, k_0 \to \infty, a = -1/2, b = 0$ (matching prior):

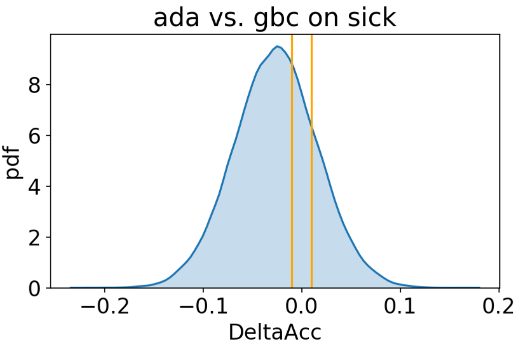
$$\mu \sim St\left(n-1, \bar{\boldsymbol{x}}, \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\rho}{1-\rho}\right)}\right)$$

- What is the difference then?

# Example

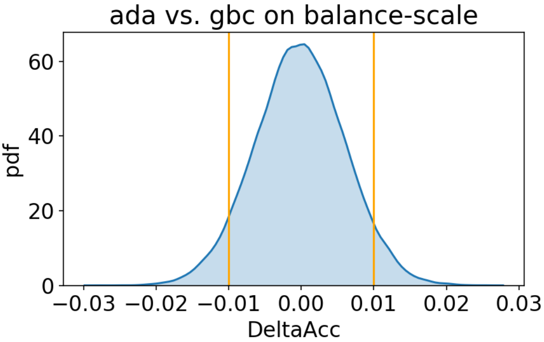
Region of practical equivalence (rope): $0.01$

$$P(ada) > gbc) = 0.65 \quad P(rope) = 0.15 \quad P(gbc > ada) = 0.20$$



# Example

- Can show practically significant differences $(1 - P(rope))$
- Can quantify uncertainty (high density intervals)
- Posterior probability of the null: $P(rope)$
- Provides basis for decisions (expected loss minimization)



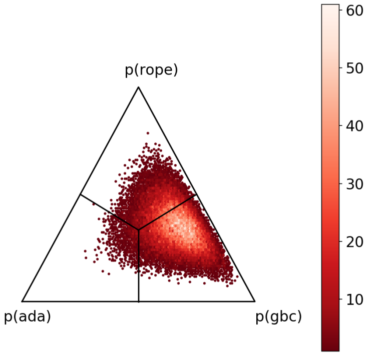
# Bayesian signed-rank test

- Let $\boldsymbol{z} = (z_1, \dots, z_q)$ be i.i.d. samples of $z$
- Place Dirichlet process prior on $z$ parameterized by strength $s > 0$ and mean $z_0$
- The posterior is a Dirichlet mixture
- Can be reformulated to a ternary distribution of test outcomes
- Monte Carlo sampling used to approximate the posterior

# Example

Rope $= 0.01$

$$P(ada > gbc) = 0.02 \quad P(rope) = 0.24 \quad P(gbc > ada) = 0.75$$




Posterior for Bayesian signed-rank test for ada vs. gbc on 27 UCI data sets

# Conclusion

- NHST has many drawbacks
- Bayesian tests:
  - claimed significant differences are practical
  - are able to detect practical equivalence
  - provide estimate with uncertainty
  - allow to automatize decisions

# Bibliography

- A. Benavoli, G. Corani, J. Demsar, and M. Zaffalon, *Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis*, ArXiv e-prints (2016).
- J. Demšar, *Statistical comparisons of classifiers over multiple data sets*, J. Mach. Learn. Res. **7** (2006), 1–30.

Thank you!

repicky at cs.cas.cz