Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 31 - Two-sample tests of the mean and applications to classifier comparison

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it
Tests and confidence intervals for classifier performance

The Caret package

1. Define sets of model parameter values to evaluate
2. for each parameter set do
   3. for each resampling iteration do
      4. Hold-out specific samples
      5. [Optional] Pre-process the data
      6. Fit the model on the remainder
      7. Predict the hold-out samples
   8. end
9. Calculate the average performance across hold-out predictions
10. end
11. Determine the optimal parameter set
12. Fit the final model to all the training data using the optimal parameter set

See R script
Binary classifier performance metrics

**Confusion matrix** (in R packages, it is transposed)

<table>
<thead>
<tr>
<th>Actual condition</th>
<th>Predicted condition</th>
<th>Information</th>
<th>Prevalence threshold (PT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive (P)</td>
<td>True positive (TP), hit</td>
<td>True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power</td>
<td>$\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$</td>
</tr>
<tr>
<td>Negative (N)</td>
<td>False positive (FP), type I error, false alarm, overestimation</td>
<td>False negative rate (FNR), miss rate</td>
<td>$\frac{FN}{P}$</td>
</tr>
<tr>
<td>Prevalence</td>
<td>Positive predictive value (PPV), precision</td>
<td>Positive likelihood ratio (LR+)</td>
<td>$\frac{TPR}{FPR}$</td>
</tr>
<tr>
<td>Accuracy (ACC)</td>
<td>False discovery rate (FDR)</td>
<td>Negative predictive value (NPV)</td>
<td>$\frac{NPV}{FN}$</td>
</tr>
<tr>
<td>Balanced accuracy (BA)</td>
<td>$\frac{F_1 \text{ score}}{2}$</td>
<td>Markedness (MK), deltaP (Δp)</td>
<td>$\frac{PPV + NPV - 1}{PPV + NPV}$</td>
</tr>
<tr>
<td></td>
<td>Fowlkes–Mallows index (FM)</td>
<td>Matthews correlation coefficient (MCC)</td>
<td>$\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$</td>
</tr>
<tr>
<td></td>
<td>Threat score (TS), critical success index (CSI), Jaccard index</td>
<td>Threat score (TS), critical success index (CSI), Jaccard index</td>
<td>$\frac{TP}{TP + FN + FP}$</td>
</tr>
</tbody>
</table>

Metrics computed on a test set are intended to estimate some parameter over the general distribution.

- $X = (W, C) \sim F$, i.e., $F$ is the (unknown) multivariate distribution of predictive features and class
- Accuracy $ACC$ of a classifier $y_\theta^+$ is a point estimate of $EF[\mathbb{1}_{y_\theta^+(W) = C}] = P_F(y_\theta^+(W) = C)$
Binary classifier performance metrics

- Binary classifier score $s_\theta(w) \in [0, 1]$ where $s_\theta(w)$ estimates $\eta(w) = P_{\theta_{\text{true}}}(C = 1|W = w)$
- ROC Curve
  - $TPR(p) = P(s_\theta(w) \geq p|C = 1)$ and $FPR(p) = P(s_\theta(w)|C = 0)$
  - ROC Curve is the scatter plot $TPR(p)$ over $FPR(p)$ for $p$ ranging from 1 down to 0
  - AUC-ROC is the area below the curve

What does AUC-ROC estimate?

- Squared error loss or $L_2$ loss or Brier score: $\frac{1}{n} \sum_i (s_\theta(w_i) - c_i)^2$
- Classifier is calibrated if $P(C = 1|s_\theta(w) = p) = p$

Binary Expected Calibration Error (binary-ECE):

- $\sum_b \frac{|B_b|}{n} |Y_b - S_b|$
  - $B_b$ is the set of $i$’s in the $b^{th}$ bin, $Y_b = |\{i| i \in B_b, c_i = 1\}|/|B_b|$, $S_b = (\sum_{i \in B_b} s_\theta(w_i))/|B_b|$
Two sample test of the mean

- Dataset \( x_1, \ldots, x_n \) realization of \( X_1, \ldots, X_n \sim F_1 \) with \( E[X_i] = \mu_1 \) and \( \text{Var}(X_i) = \sigma^2_X \)
- Dataset \( y_1, \ldots, y_m \) realization of \( Y_1, \ldots, Y_m \sim F_2 \) with \( E[Y_i] = \mu_2 \) and \( \text{Var}(Y_i) = \sigma^2_Y \)
  - measurements for control and (medical) treatment groups of patients
  - performances on benchmark datasets/folds of two different classifiers
- \( H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2 \)
- Wald test statistics: \( T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\text{Var}(\bar{X}_n - \bar{Y}_m)}} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}}} \)

We distinguish a few cases:
- \( F_1, F_2 \) are normal distributions
  - \( \sigma^2_X \) and \( \sigma^2_Y \) are known [z-test]
  - \( \sigma^2_X \) and \( \sigma^2_Y \) are unknown and \( \sigma^2_X = \sigma^2_Y \) [t-test]
  - \( \sigma^2_X \) and \( \sigma^2_Y \) are unknown and \( \sigma^2_X \neq \sigma^2_Y \) [Welch test]
- \( F_1, F_2 \) are general distributions
  - Large sample [t-test]
  - \( F_1(x - \Delta) = F_2(x) \) location-shift
  - Bootstrap two sample test
- Paired data [paired t-test]
Normal data with known $\sigma^2_X$ and $\sigma^2_Y$: z-test

- $X_1, \ldots, X_n \sim N(\mu_1, \sigma^2_X)$ and $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma^2_Y)$

- $H_0: \mu_1 = \mu_2$

- $H_1: \mu_1 \neq \mu_2$

- $100(1 - \alpha)\%$, e.g., 95% or 99% or 99.9%
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$

- $Z = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}}} \sim N(0, 1)$ test statistics when $H_0$ is true

- $z$ value is $\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}}}$ and $p$-value $p = P(|Z| \geq |z|) = 2(1 - \Phi(|z|))$

- $P(Z \leq -z_{\alpha/2}) = \alpha/2$ and $P(Z \geq z_{\alpha/2}) = \alpha/2$

- Output of the test at confidence level $100(1 - \alpha)\%$ using critical values
  - $|z| \geq z_{\alpha/2}$: $H_0$ is rejected
  - otherwise: $H_0$ cannot be rejected

See R script
Unknown $\sigma^2_X = \sigma^2_Y = \sigma^2$ and pooled variance

- We need to estimate $\text{Var}(\bar{X}_n - \bar{Y}_m) = \sigma^2 \left( \frac{1}{n} + \frac{1}{m} \right)$

- Recall

  $$S^2_X = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \quad \text{and} \quad S^2_Y = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \bar{Y}_m)^2$$

  are unbiased estimators of $\sigma^2_X$ and $\sigma^2_Y$

- The pooled variance:

  $$S^2_p = \frac{(n-1)S^2_X + (m-1)S^2_Y}{n+m-2} \left( \frac{1}{n} + \frac{1}{m} \right) = \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 + \sum_{i=1}^{m} (Y_i - \bar{Y}_m)^2}{n+m-2} \left( \frac{1}{n} + \frac{1}{m} \right)$$

  is an unbiased estimator of $\sigma^2 \left( \frac{1}{n} + \frac{1}{m} \right)$
Testing equal variances for normal data: \( F \)-test

- \( X_1, \ldots, X_n \sim N(\mu_1, \sigma_X^2) \) and \( Y_1, \ldots, Y_m \sim N(\mu_2, \sigma_Y^2) \)
- \( H_0 : \sigma_X^2 = \sigma_Y^2 \)
- \( H_1 : \sigma_X^2 \neq \sigma_Y^2 \)  
  \[ \text{[Two-tailed test]} \]
- \( 100(1 - \alpha)\% \), e.g., 95\% or 99\% or 99.9\%  
  \[ \text{i.e., } \alpha = 0.05 \text{ or } \alpha = 0.01 \text{ or } \alpha = 0.001 \]  
  \[ \text{[Confidence level]} \]
- \( F = \frac{s_X^2}{s_Y^2} \sim F(n - 1, m - 1) \) test statistics when \( H_0 \) is true  
  \[ \text{[Fisher-Snedecor distribution]} \]
- \( f \) value is \( \frac{s_X^2}{s_Y^2} \) and \( p \)-value is \( p = 2 \min \{ P(F \leq f), 1 - P(F \leq f) \} \)  
  \[ \text{[Asymmetric]} \]
- \( P(F \leq l) = \alpha/2 \) and \( P(F \geq u) = \alpha/2 \)  
  \[ \text{[Critical values]} \]
- Output of the test at confidence level \( 100(1 - \alpha)\% \) using critical values  
  \[ \text{[Critical region]} \]
  - \( f \leq l \) or \( f \geq u \) : \( H_0 \) is rejected  
  - otherwise: \( H_0 \) cannot be rejected

See R script
Common distributions

- Probability distributions at Wikipedia
- Probability distributions in R
Normal data with unknown $\sigma^2_X = \sigma^2_Y = \sigma^2$: t-test

- $X_1, \ldots, X_n \sim N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma^2)$
- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$ [Two-tailed test]
- $100(1 - \alpha)\%$, e.g., 95% or 99% or 99.9%
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ [Confidence level]
- $T_p = \frac{\bar{X}_n - \bar{Y}_m}{S_p} \sim t(n + m - 2)$ test statistics when $H_0$ is true
- $t$ value is $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{(n-1)s^2_X + (m-1)s^2_Y}{n+m-2} \left(\frac{1}{n} + \frac{1}{m}\right)}}$ and $p$-value $p = P(|T_p| \geq |t|)$ [Significance level]
- $P(T_p \leq -t_{n+m-2,\alpha/2}) = \alpha/2$ and $P(T_p \geq t_{n+m-2,\alpha/2}) = \alpha/2$ [Critical values]
- Output of the test at confidence level $100(1 - \alpha)\%$ using critical values
  - $|t| \geq t_{n+m-2,\alpha/2}$: $H_0$ is rejected [Critical region]
  - otherwise: $H_0$ cannot be rejected

See R script
Normal data with unknown $\sigma^2_X \neq \sigma^2_Y$

• The nonpooled variance:

$$S_d^2 = \frac{S_X^2}{n} + \frac{S_Y^2}{m}$$

is an unbiased estimator of $\text{Var}(\bar{X}_n - \bar{Y}_m) = \frac{\sigma^2_X}{n} + \frac{\sigma^2_Y}{m}$

• The test statistics $T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d}$ is not $t$-distributed!

• Possible solution: use methods for general distributions (see later slides)

• Another solution: Welch t-test (next slide)
Normal data with unknown $\sigma_X^2 \neq \sigma_Y^2$: Welch t-test

- $X_1, \ldots, X_n \sim N(\mu_1, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma_Y^2)$
- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$ [Two-tailed test]
- $100(1 - \alpha)\%$, e.g., 95\% or 99\% or 99.9\%
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ [Confidence level]
- $T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d} \approx t(\nu)$ test statistics when $H_0$ is true, with $\nu = \frac{\left(\frac{1}{n} + \frac{u}{m}\right)^2}{\frac{1}{n^2(n-1)} + \frac{u^2}{m^2(m-1)}}$ and $u = \frac{s_Y^2}{s_X^2}$ [Significance level]
- $t$ value is $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$
- $p$-value $p = P(|T_d| \geq |t|)$
- $P(T_d \leq -t_{\nu,\alpha/2}) = \alpha/2$ and $P(T_d \geq t_{\nu,\alpha/2}) = \alpha/2$ [Critical values]
- Output of the test at confidence level $100(1 - \alpha)\%$ using critical values
  - $|t| \geq t_{\nu,\alpha/2}$: $H_0$ is rejected [Critical region]
  - otherwise: $H_0$ cannot be rejected

See R script
General data, large sample: t-test

- $X_1, \ldots, X_n \sim F_1$ and $Y_1, \ldots, Y_m \sim F_2$
- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

- 100$(1 - \alpha)$%, e.g., 95% or 99% or 99.9%
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$

- $T_d = \frac{\bar{x}_n - \bar{y}_m}{S_d} \approx N(0, 1)$
- $t$ value is $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$ and $p$-value $p = P(|T_d| \geq |t|)$
- $P(T_d \leq -z_{\alpha/2}) = \alpha/2$ and $P(T_d \geq z_{\alpha/2}) = \alpha/2$

- Output of the test at confidence level 100$(1 - \alpha)$% using critical values
  - $|t| \geq z_{\alpha/2}$: $H_0$ is rejected
  - otherwise: $H_0$ cannot be rejected

See R script
General data, location-shift: Wilcoxon rank-sum test

- Also called as: **Mann–Whitney U test** or Mann–Whitney–Wilcoxon (MWW)
- \(X_1, \ldots, X_n \sim F_1\) and \(Y_1, \ldots, Y_m \sim F_2\)
- \(H_0: \mu_1 = \mu_2\) and \(H_1: \mu_1 \neq \mu_2\) [Two-tailed test]
  - actually, \(H_0: F_1(x - \Delta) = F_2(x)\) where \(\Delta = \mu_2 - \mu_1\) [Location-shift model]
  - we should test that empirical distributions have the same shape
- \(W = \sum_{i=1}^{n} S_i \sim W(n, m)\) when \(H_0\) is true [or \(U = W - m \cdot (m + 1)/2\)]
  - where \(S_i\) is the rank of \(X_i\) in sorted\((X_1, \ldots, X_n, Y_1, \ldots, Y_m)\)
  - \(\text{pwilcox}\) in R, or large sample Normal approx
- \(w\) value is \(\sum_{i=1}^{n} s_i\) and \(p\)-value \(p = P(|W| \geq |w|)\)
- \(P(W \leq -w_{\alpha/2}) = \alpha/2\) and \(P(T_P \geq w_{\alpha/2}) = \alpha/2\) [Critical values]
- Output of the test at confidence level 100\((1 - \alpha)\)% using critical values [Critical region]
  - \(|w| \geq w_{\alpha/2}: H_0\) is rejected
  - otherwise: \(H_0\) cannot be rejected
- Generalized test (without location-shift assumption): **Brunner-Munzel** test

See R script
General data: bootstrap test

- Equal variance ($\sigma^2_X = \sigma^2_Y$)
  - bootstrap of pooled studentized mean difference
    \[ t^*_p = \frac{(\bar{x}^*_n - \bar{y}^*_m) - (\bar{x}_n - \bar{y}_m)}{s^*_p} \]

- Non-equal variance ($\sigma^2_X \neq \sigma^2_Y$)
  - bootstrap of nonpooled studentized mean difference
    \[ t^*_d = \frac{(\bar{x}^*_n - \bar{y}^*_m) - (\bar{x}_n - \bar{y}_m)}{s^*_d} \]

See R script
Paired data

- Datasets $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are measurements for the same experimental unit
  - unit: a person before and after a (medical) treatment
  - unit: a dataset/fold used to train two different classifiers
- The theory is essentially based on taking differences $x_1 - y_1, \ldots, x_n - y_n$ and thus reducing the problem to that of a one-sample test.
- $H_0 : \mu_1 = \mu_2 \Rightarrow H_0 : \mu_1 - \mu_2 = 0$
- Advantage: better power / lower Type II risk of the test w.r.t. unpaired version
  - $P_{\text{paired}}(p \leq \alpha | H_1) \geq P_{\text{unpaired}}(p \leq \alpha | H_1)$

See R script
Optional reference

- On confidence intervals and statistical tests (with R code)

Myles Hollander, Douglas A. Wolfe, and Eric Chicken (2014)
Nonparametric Statistical Methods.