Assessing Models

Performance

\[ \begin{align*}
\rightarrow & \quad \text{Precision} \\
\rightarrow & \quad \text{Recall} \quad \text{per class basis} \\
\rightarrow & \quad \text{F-score}
\end{align*} \]

Binary classification \( (P, N) \)

Contingency

True

Prediction

\[ \begin{align*}
\text{test} \quad x_i \rightarrow \hat{y}_i \quad \text{predicted} \\
\text{vs} \quad y_i \quad \text{true}
\end{align*} \]

\[ \begin{align*}
\text{TP: } y_i = P, \; \hat{y}_i = P & \quad \text{True positive} \\
\text{TN: } y_i = N, \; \hat{y}_i = N & \quad \text{True negative} \\
\text{FP: } y_i = N, \; \hat{y}_i = P & \quad \text{False positive} \\
\text{FN: } y_i = P, \; \hat{y}_i = N & \quad \text{False negative}
\end{align*} \]

overall Error rate: \[ \frac{FP + FN}{N} \]

Accuracy: \[ \frac{TP + TN}{N} \]

Precision \( P \):
\[ \frac{TP}{n_1} = \frac{TP}{(TP + FP)} \]

\( N \):
\[ \frac{TN}{n_2} = \frac{TN}{(TN + FN)} \]

ROC Analysis

Sensitivity \( TPR \)

Recall \( P \):
\[ \frac{TP}{n_1} = \frac{TP}{(TP + FN)} \]

\( N \):
\[ \frac{TP}{n_1} = \frac{TP}{(TP + FN)} \]

\( \equiv \) sensitivity

\( TPR \): True Positive Rate

\[ = \text{recall} (P) \]

\[ = \frac{TP}{TP + FN} \]

\[ \equiv \text{sensitivity} \]

\( FPR = 1 - TNR \)
Binary classifier: Naïve Bayes

\[ z \rightarrow p(p|z) / p(N|z) \]

\( \theta \): default threshold for \( p = 0.5 \)

Test set: \( n \) points

Sort the points in decreasing order of \( p(p|z) \)

M

\[ x_5, x_4, x_3, x_2, x_1 \]

0.9 0.8 0.8 0.3 0.1

\[ \begin{array}{cccc}
\text{P} & \text{P} & \text{P} & \text{P} \\
\text{P} & \text{N} & \text{N} & \text{N} \\
\text{P} & \text{N} & \text{N} & \text{N} \\
\text{P} & \text{N} & \text{N} & \text{N} \\
\end{array} \]

\( \theta = 1.0 \)

\( \theta = 0.9 \)

\( \theta = 0.8 \)

\( \theta = 0.3 \)

ROC curve

TP\( / (TP + FN) = TPR \)

Correct: vertical up

Incorrect: horizontal right

Perfect classifier

Classifier A (Naïve Bayes)

Classifier B (SVM)

\[ \begin{array}{ccc}
\theta & TPR & FPR \\
1.0 & 0.0 & 0.0 \\
0.9 & 0.0 & 0.0 \\
0.1 & 1.0 & 1.0 \\
\end{array} \]

Area under the curve (Auc)

Choose the classifier with the higher AUC

AUC for ideal classifier = 1.0
Auc for random classifier = 0.5

Cross-validation & Comparison of models (Roc/Auc)

Significance testing (test is also a random sample)

K-fold CV: cross validation

1 fold

D1: test

D2: train

Dk: test

D1...Dk: training

M1

Mk

θi: the performance of classifier Mi

θ̄ = \frac{1}{k} \sum θi ← avg performance

σ²₂ = \frac{1}{k} \sum (θi - θ̄)^2 ← variance of performance

σθ = std

K: 5

5-fold
test: 20%

10-fold
test: 10%

When data is too small

R...
When data is too small

**Bootstrap Validation / Resampling**

- Draw a size $n$ sample with replacement from $D$
- $D_1 \rightarrow M_1 \rightarrow \Theta_1$
- $D_2 \rightarrow M_2 \rightarrow \Theta_2$
- $D_k \rightarrow M_k \rightarrow \Theta_k$

Optimistic / biased

$\Theta \leftarrow \text{mean}$

$\Theta^2 \leftarrow \text{variance}$

$D_i$ (Bootstrap sample): $\hat{\Theta}_i \in D$

- What is the probability that $\hat{\Theta} \notin D_i$?

Bernoulli

\[
\left[ \frac{1}{n} = p \\
\text{prob that } \hat{\Theta} \text{ is not chosen} \right]
\]

\[
1 - \frac{1}{n} = q = 1 - p
\]

\[
\text{prob that } \hat{\Theta} \text{ is not chosen even after } n \text{ trials}
\]

\[
E^n = \left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e} = 0.368
\]

The overlap between $D$ & $D_i$ is 63.2%.

**Confidence Intervals & Significance testing**
**Confidence Intervals & Significance Testing**

**K-fold CV**

\[ D \rightarrow D/D_1 \rightarrow M_1 \rightarrow \Theta_1 (D_1) \]

\[ D/D_2 \rightarrow M_2 \rightarrow \Theta_2 (D_2) \]

\[ \vdots \]

\[ D/D_K \rightarrow M_K \rightarrow \Theta_K (D_K) \]

**Train**

**Measure**

\[ \widehat{\mu}_0 = \text{estimated mean} \]

\[ \sigma_0^2 = \text{estimated variance} \]

from \( k \) samples / folds

**True mean performance, e.g., error rate?**

\[ \mu_0 = \]  

\[ P \left( \mu_0 - 5 \sigma < \mu_0 < \mu_0 + 5 \sigma \right) \approx 95\% \]

**Confidence interval**

ideally: small interval

\[ \Theta_1, \Theta_2, \ldots, \Theta_K \]

\[ \widehat{\mu}_0 \leftarrow \text{Random variable} \]

Sample size \( K \)

\[ E[\widehat{\mu}_0] = \mu_0 \]

Unbiased estimator

\[ \text{Var}[\widehat{\mu}_0] = \frac{\sigma_0^2}{K} \]

\[ Z_k = \frac{\widehat{\mu}_0 - E[\widehat{\mu}_0]}{\text{std}(\widehat{\mu}_0)} \]

\[ Z_k = \frac{\widehat{\mu}_0 - \mu_0}{\sigma_0/\sqrt{K}} \]

\( Z_k \) follows a normal distribution \( Z_k \sim N(0, 1) \)

as \( K \to \infty \)
As \( K \to \infty \)

\[ \frac{-z_{0.12}}{\sigma} \leq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{K}} \leq \frac{z_{0.12}}{\sigma} \]

\[ P \left( -z_{0.12} \leq z_K \leq z_{0.12} \right) = 1 - \alpha \]

\[ -z_{0.12} \cdot \frac{\sigma}{\sqrt{K}} \leq \overline{x} - \mu_0 \leq z_{0.12} \cdot \frac{\sigma}{\sqrt{K}} \]

\[ P \left( \frac{\overline{x} - \mu_0}{\sigma/\sqrt{K}} \right) = 1 - \alpha \]

Confidence 95%

For large sample, \( z_{0.12} \) from Normal

For small samples: \( t \)-distribution instead of Normal

\[ k = 5 \]

\[ \text{df} \]

\[ \mu_0 - \frac{z_{0.12}}{\sqrt{k}} \leq \overline{x} \leq \mu_0 + \frac{z_{0.12}}{\sqrt{k}} \]
Significance testing

Compare 2 models
\( M_A \) and \( M_B \)

CV setting

Paired t-test

Q: Is the performance of \( M_A \) significantly different than \( M_B \)?

\[
\begin{array}{c|c}
A & B \\
\hline
\Theta_1^A & \Theta_1^B \\
\Theta_2^A & \Theta_2^B \\
\vdots & \vdots \\
\Theta_k^A & \Theta_k^B \\
\end{array}
\]

\[ \text{diff} = s_1 - s_1^2 \ldots s_k \]

\[ s_k = \Theta_k^A - \Theta_k^B \]

\[ \hat{M}_s = \frac{\sum s_k}{k} \]

Hypothesis test:

- Null hypothesis: \( H_0 \equiv M_s = 0 \)
- Alternative hypothesis: \( H_A \equiv M_s \neq 0 \)

Accept if \( H_0 \) is rejected.
Under null hypothesis $\mu_S = 0$

$$Z_K = \frac{\hat{\mu}_S - \mu_S}{\hat{\sigma}_S / \sqrt{k}} = \frac{\sqrt{k}}{\hat{\sigma}_S}$$

$t$-distribution

$t$-distribution

$(k-1$ dof $)$

Two-tailed $t$-test paired

$$\alpha = 5\%$$

Confidence = 95%

If $Z_K \in [-t_{1/2}, t_{1/2}]$ then accept the null hypothesis

**Rejection region (rejecting $H_0$)**

If $Z_K < -t_{1/2}$ or $Z_K > t_{1/2}$ then $H_0$ can be rejected at 95%.

$$(1-\alpha)$$

$H_0$: $A$ and $B$ are no different

$$\alpha = 5\%$$
Null Hypothesis: $H_0: \theta^A \leq \theta^B$

One-tailed test

Supervised methods: there is no distinguished attribute like response variable

Unsupervised methods: find "natural" groups

$V \in \mathbb{R}^d$