Evaluating Classifiers

Volker Tresp
Summer 2015
How Useful is a Classifier?

- We have trained a classifier. For a given input $x$ the classifier either predicts a 0 or a 1. If the classifier produces a score (e.g., a posterior probability), we apply a threshold, such that, again, a 0 or a 1 is produced as output.

- How useful is a particular classifier in different scenarios?

- The quantity of interest is

$$P(y, \text{pred} = j) = \int P(x, y) I(f(x) = j) dx$$

where $f(x) \in \{0, 1\}$ is the output of the classifier and also $j \in \{0, 1\}$. $I(\cdot)$ is the indicator function.

- Note that $P(x, y)$ might or might not reflect the distribution under which the classifier was trained, i.e., $\bar{P}(x, y)$.
Invariances

- Causal Model I: If there is an underlying causal model where $y$ is a cause of $x$, then one assumes that $P(x|y) = \tilde{P}(x|y)$ is stable in any experiment. Example: $y$ is a fire, $x$ is sensory information to a fire alarm and $f(x)$ is the fire alarm. Then the probability of a fire $P(y)$ might be different in different buildings but $P(x|y)$ is identical in all buildings.

- Causal Model II: If there is an underlying causal model where $x$ is a cause of $y$, then one assumes that $P(y|x) = \tilde{P}(y|x)$ is stable in any experiment. Example: $x$ is age and $y$ is cancer. In different cities $P(x)$ might differ, but $P(y|x)$ is identical in all cities.
Empirical Estimates

• One approximates

\[ P(\text{pred} = i, y = j) \approx \frac{N_{i,j}}{N} \]

where the data represent the test distribution. \( N \) is the total number of observations in the test set.

• TP stands for \textit{true positive} or \textit{hit} and is defined as

\[ TP = N_{\text{true, true}} \]

• TP stands for \textit{true negative} or \textit{correct rejection} and is defined as

\[ TN = N_{\text{false, false}} \]

• FP stands for \textit{false positive}, \textit{false alarm} or \textit{Type I error} and is defined as

\[ FP = N_{\text{true, false}} \]
• FN stands for \textit{false negative}, \textit{miss} or \textit{Type II error} and is defined as

\[ FN = N_{false, true} \]
Common Performance Measures

- Although these numbers tell the story one often calculates additional indicators. For example one might be interested in the percentage of fires that are detected.

\[ P(pred = 1|y = 1) = \frac{TP}{TP + FN} = \text{Recall} \]

Recall is also called sensitivity, true positive rate, hit rate, or detection rate.

- Or one might be interested in how often an alarm is released, when there really is a fire.

\[ P(y = 1|pred = 1) = \frac{TP}{TP + FN} = \text{Precision} \]

Precision is also called positive predicted value.

- Another quantity is the

\[ P(pred = 0|y = 0) = \frac{TN}{TN + FP} = \text{Specificity} \]

Specificity is also called true negative rate.
And there is

\[
P(y = 0|pred = 0) = \frac{TN}{TN + FN} = \text{Negative Predicted Value}
\]
Invariances

• Note that Recall and Specificity are invariant to $P(y)$ and under Causal Model I reflect the properties of the detector

$$P(pred = j|y) = \frac{1}{P(y)} \int P(y) \tilde{P}(x|y) I(f(x) = j) dx$$

$$= \int \tilde{P}(x|y) I(f(x) = j) dx$$

• Note that Precision and Negative Predicted Value are invariant to $P(x)$ and under Causal Model II reflect the properties of the detector

$$P(y|pred = j) = \frac{1}{P(pred = j)} \int P(x) \tilde{P}(y|x) I(f(x) = j) dx$$

$$= \int \tilde{P}(y|x) I(f(x) = j) dx$$
Definitions
A 2x2 confusion matrix is shown with the following labels:

- True Positive (TP)
- False Negative (FN)
- True Negative (TN)
- False Positive (FP)

The matrix is divided into four quadrants, with the axes showing the prediction (\( \hat{y} \)) on the horizontal axis and the true outcome (\( y \)) on the vertical axis.
Running Example:

\[ \text{pred} = \hat{y} \]

\[
\begin{array}{c|c}
 \text{pred} & \text{true} \\
\hline
1 & FN (0.02, 20) \\
0 & TN (0.3, 940) \\
\hline
1 & TP (0.04, 30) \\
0 & FP (0.01, 10) \\
\end{array}
\]
Probabilistic Interpretation

- with $N = TP + FP + TN + FN$ test patterns,

  \[
  \hat{P}(pred = 1, y = 1) = \frac{TP}{N}
  \]

  \[
  \hat{P}(pred = 1, y = 0) = \frac{FP}{N}
  \]

  \[
  \hat{P}(pred = 0, y = 0) = \frac{TN}{N}
  \]

  \[
  \hat{P}(pred = 0, y = 1) = \frac{FN}{N}
  \]
Accuracy

- **Accuracy**: 
  \[ \text{Accuracy} = \frac{TP + TN}{N} \]

- If we assign the label *correct* to the events \((\text{pred} = 1, y = 1)\) and \((\text{pred} = 0, y = 0)\), then 
  \[ \text{Accuracy} = P(\text{correct}) \]

- The *error rate* is \(1 - \text{Accuracy}\).

- Accuracy is not a useful measure for highly imbalanced classes where trivial classifiers (always predict 0 or 1 independent of input) can already have high accuracy but are useless.

- In the running example: \(\text{Accuracy} = 0.97\) and the error rate is 0.03
**Precision**

- **Precision** (Relevance). Also called positive predicted value (PPV)

\[
Precision = \frac{TP}{TP + FP}
\]

- “What’s the percentage of good fish in my catch”
- This approximates

\[
P(y = 1|pred = 1)
\]

- In our running example, precision is 0.75
Recall

- **Recall** (sensitivity, true positive rate, hit rate, detection rate):

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

- “How many good fish did I catch if compared to all fish in the ocean”

- This approximates

\[
P(pred = 1|y = 1)
\]

- In our running example, recall is 0.60
Specificity

- **Specificity** (true negative rate, 1 - false-positive-rate, 1-false alarm rate)

\[
Specificity = \frac{TN}{TN + FP}
\]

- This approximates

\[
P(pred = 0 | y = 0)
\]

- In our running example specificity is 0.98
Negative Predictive Value

• **Negative Predictive Value** (NPV)

\[ NPV = \frac{TN}{TN + FN} \]

• This approximates

\[ P(y = 0 | pred = 0) \]

• Not relevant for search engines since, even for lousy search engines, close to one

• PPV (precision) and NPV are used by doctors to evaluate the consequences of test results for a particular patient

• In our running example NPV is 0.97
F-Measure

• F-measure

\[ F = 2 \frac{Precision \times Recall}{Precision + Recall} \]

The F-measure combines precision and recall. Trivial search engines, that either predict all pages to be relevant or irrelevant, would have an F-measure of 0.

• In our running example the F-measure is 0.66
Odds and Odds Ratio

• We can interpret the treatment as $pred$ and outcome as $y$

• Then

$$(Odds|treatment = 1) = \frac{TP}{FP}$$

$$(Odds|treatment = 0) = \frac{FN}{TN}$$

• The odds ratio then is

$$OR = \frac{TP \times TN}{FP \times FN} = \frac{P(y = 1|pred = 1)P(y = 0|pred = 0)}{P(y = 0|pred = 1)P(y = 1|pred = 0)}$$

$$= \frac{P(pred = 1|y = 1)P(pred = 0|y = 0)}{P(pred = 0|y = 1)P(pred = 1|y = 0)}$$

• The OR is stable both under Causal Model I and Causal Model II

• In the running example $OR = 141$
Rankings and Cut-off

- Most classifiers do not just produce a decision (0/1) but also a ranking.
- For most classifiers we can define a variable discrimination threshold which determines which patterns are classified as ones and zeros.
### Evaluation on test set

- **TP (True Positive=Hit)** = \#tp (here: 3) [inner product]
- **FP (False Positive=False Alarm=type I error)** = \#fp (here: 2)
- **TN (True Negative)** = \#tn (here: 2)
- **FN (False Negative=Miss=type II error)** = \#fn (here: 1)
In the ROC (Receiver operating characteristic) curve, one varies $\alpha$ and plots Recall (y-axis) against (1-Specificity = FPR) (x-axis).

Advantage: The ROC is independent of the class mix and purely reflects the performance of the classifier!

To obtain an overall measure of classification quality one forms the integral under the curve and obtains the AUC-ROC. A random classifier has an AUC-ROC of 0.5, a perfect classifier of 1.

AUC-ROC can be shown to be equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.
Die Receiver Operating Characteristic (ROC) - Kurve

- Gibt mein Klassifikator eine Klassenwahrscheinlichkeit aus, dann entscheide ich mich für Klasse 0, wenn dieser Wert unter einem Schwellewert S ist und ansonsten entscheide ich mich für Klasse 1.
- $(0,0): S=1 \ (\alpha=-\infty)$  
  $(1,1): S=0 \ (\alpha=\infty)$  
  $(0,3, 0.85): S=0.5 \text{ (Beispiel)}$

- Ich sage nur Klasse 1 voraus
- Mein Klassifikator sagt, dass jede Webseite ein Treffer ist
- Alle Patienten sind krank (Klasse 1)
- Spezifität: $P(\text{pred} = 0 \mid y = 0) = 0$  
  Sensitivität: $P(\text{pred} = 1 \mid y = 1) = 1$

- Ich sage nur Klasse 0 voraus
- Mein Klassifikator sagt, dass keine Webseite ein Treffer ist
- Alle Patienten sind gesund (Klasse 0)
- Spezifität: $P(\text{pred} = 0 \mid y = 0) = 1$  
  Sensitivität: $P(\text{pred} = 1 \mid y = 1) = 0$

- Das Integral unter der Kurve (area under curve, AUC-ROC) ist bei perfekter Klassifikation gleich 1 und bei Zufallsklassifikation gleich 0.5
PR-Curve and AUC-PR

- For a search engine precision and recall are important
- In the PR curve on plots precision (y-axis) against recall (x-axis)
- To obtain an overall measure of classification quality one forms the integral under the curve and obtains the AUC-PR. A perfect classifier has an AUC-PR of 1
Evaluating Search Engines

- AUC-PR is a good measure for the evaluation of a search engine

- nDCG (normalized discounted cumulative gain) is also often used to evaluate search engines. One gets a high score if the highest ranked hits have a large relevant score. nDCG is insensitive to ranking mistakes at lower ranked positions