Statistics for Data Science
Lesson 03 - Bayes’ rule and applications

Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
salvatore.ruggieri@unipi.it
Exercise at home from Lesson 01

Exercise at home. Prove or disprove:

- If $A$ is independent of $B$ then $A$ is conditionally independent of $B$ given $C$

  In formula, if $P(A \cap B) = P(A)P(B)$ then $P(A \cap B|C) = P(A|C)P(B|C)$

Counterexample.

- $\Omega = \{H, T\} \times \{H, T\}$ two coin toss
- $A = \{\text{first coin is H}\} = \{(H, H), (H, T)\}$ $P(A) = 1/2$
- $B = \{\text{second coin is H}\} = \{(H, H), (T, H)\}$ $P(B) = 1/2$

\[
P(A \cap B) = 1/4 = P(A)P(B)
\]

- $C = \{\text{both coins have same result}\} = \{(H, H), (T, T)\}$ $P(C) = 1/2$

\[
P(A \cap B|C) = \frac{P(A \cap B \cap C)}{P(C)} = 1/2 \neq P(A|C)P(B|C) = \frac{P(A \cap C)P(B \cap C)}{P(C)^2} = 1/4
\]

Same counterexample shows that pairwise independence is weaker than independence: $A, B, C$ are pairwise independent, but not independent!
Exercise

Exercise. Prove or disprove:

- If $A$, $B$ and $C$ are independent, then $A$ is conditionally independent of $B$ given $C$ (i.e., $P(A \cap B | C) = P(A | C)P(B | C)$)

Proof. Independence implies $P(A \cap B \cap C) = P(A)P(B)P(C)$ and then:

$$P(A \cap B | C) = \frac{P(A \cap B \cap C)}{P(C)} = \frac{P(A)P(B)P(C)}{P(C)} = P(A)P(B)$$

Independence also implies $P(A \cap C) = P(A)P(C)$ and $P(B \cap C) = P(B)P(C)$, and then:

$$P(A | C)P(B | C) = \frac{P(A \cap C)P(B \cap C)}{P(C)^2} = \frac{P(A)P(C)P(B)P(C)}{P(C)^2} = P(A)P(B)$$
Testing for Covid-19

A new test for Covid-19 (or Mad-Cow disease, or drug use) has been developed.

- $\Omega = \{ \text{people aged 18 or higher} \}$
- $+ = \{ \text{people tested positive} \} \quad - = \{ \text{people tested negative} \} = +^c$
- $C = \{ \text{people with Covid-19} \} \quad C^c = \{ \text{people without Covid-19} \}$

In lab experiments, a sample of people with and without Covid-19 tested

- $P(+) = 0.99$  
  \text{[Sensitivity/Recall/True Positive Rate]}
- $P(−|C^c) = 0.99$  
  \text{[Specificity/True Negative Rate]}

What is the probability I really have Covid-19 given that I tested positive?  \text{[Precision]}

\[
P(C|+) = \frac{P(C \cap +)}{P(+)} = \frac{P(+) \cdot P(C)}{P(+)} = \frac{P(+) \cdot P(C)}{P(+) \cdot P(C) + P(+|C^c) \cdot P(C^c)}
\]

\[
P(C|+) = \frac{0.99 \cdot P(C)}{0.99 \cdot P(C) + 0.01 \cdot (1 - P(C))}
\]
$P(C)$, the probability of having Covid-19, is unknown. Let’s plot $P(C|+)$ over $P(C)$:

- For $P(C) = 0.02$, $P(C|+) = 0.67$
- For $P(C) = 0.06$, $P(C|+) = 0.86$
- For $P(C) = 0.10$, $P(C|+) = 0.92$

See R script
Bayes’ Rule

It follows from \( P(C_i | A) = \frac{P(A | C_i) \cdot P(C_i)}{P(A)} \) and the law of total probability

Useful when:
- \( P(C_i | A) \) not easy to calculate
- while \( P(A | C_j) \) and \( P(C_j) \) are known for \( j = 1, \ldots, m \)
- E.g., in classification problems (see Bayesian classifiers from Data Mining)

\( P(C_i) \) is called the prior probability

\( P(A | C_i) \) is called the posterior probability (after seeing event \( C_i \))
Binary Classifiers

- $\Omega = \{f, m\} \times \mathbb{N} \times \{+, -\}$

- Features:
  - $G$ gender, $G = f$ is $\{\omega \in \Omega \mid \omega = (f, -, -)\}$
  - $A$ age, $A = 25$ is $\{\omega \in \Omega \mid \omega = (-, 25, -)\}$
  - $Y$ true class
    - $Y = +$ is $\{\omega \in \Omega \mid \omega = (-, -, +)\}$, e.g., Covid-19 positive
    - $Y = -$ is $\{\omega \in \Omega \mid \omega = (-, -, -)\}$, e.g., Covid-19 negative

- Binary Classifier: $\hat{Y} : \{f, m\} \times \mathbb{N} \rightarrow \{+, -\}$ predicted class
  - $\hat{Y} = +$ is $\{(g, a, c) \in \Omega \mid \hat{Y}((g, a)) = +\}$, e.g., predicted Covid-19 positive
  - $\hat{Y} = -$ is $\{(g, a, c) \in \Omega \mid \hat{Y}((g, a)) = -\}$, e.g., predicted Covid-19 negative

- $P(Y = \hat{Y})$, i.e., $P(Y = + \cap \hat{Y} = +) + P(Y = - \cap \hat{Y} = -)$ [True Accuracy]

- $P(Y = + | \hat{Y} = +)$ [True Precision]

- $P(\hat{Y} = + | Y = +)$ [True Recall]

- Such probabilities are unknown! They can only be estimated on a sample (test set)
Precision of classifiers

**Confusion matrix** over the test set!

<table>
<thead>
<tr>
<th></th>
<th>True $Y$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$+$</td>
<td>$-$</td>
</tr>
<tr>
<td>Predicted $\hat{Y}$</td>
<td>$+$</td>
<td>$\text{TP}$</td>
</tr>
<tr>
<td>$-$</td>
<td>$\text{FN}$</td>
<td>$\text{TN}$</td>
</tr>
<tr>
<td>Total</td>
<td>$P$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

- $P(\hat{Y} = +| Y = +) \approx \frac{\text{TP}}{P}$ \[Sensitivity/Recall/TPR\]
- $P(\hat{Y} = -| Y = -) \approx \frac{\text{TN}}{N}$ \[Specificity/TNR\]
- “$\approx$” reads as “approximatively” \[Probability estimation\]

What is the probability I really am positive given that I was predicted positive? \[Precision\]

$$P(Y = +| \hat{Y} = +) = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\frac{8}{16}$$
### Precision of classifiers

**Confusion matrix** over the test set!

<table>
<thead>
<tr>
<th></th>
<th>True Y</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>TP</td>
<td>−</td>
</tr>
<tr>
<td>−</td>
<td>FN</td>
<td>+</td>
</tr>
<tr>
<td>Total</td>
<td>PP</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td>PN</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>P</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>P + N</td>
<td></td>
</tr>
</tbody>
</table>

- $P(\hat{Y} = + | Y = +) \approx \frac{TP}{P}$ [Sensitivity/Recall/TPR]
- $P(\hat{Y} = − | Y = −) \approx \frac{TN}{N}$ [Specificity/TNR]
- “≈” reads as “approximatively” [Probability estimation]

**What is the probability I really am positive given that I was predicted positive?** [Precision]

$$P(Y = + | \hat{Y} = +) = \frac{P(\hat{Y} = + | Y = +) \cdot P(Y = +)}{P(\hat{Y} = + | Y = +) \cdot P(Y = +) + (1 - P(\hat{Y} = − | Y = −)) \cdot P(Y = −)}$$

$$\approx \frac{TP/P \cdot P(Y = +)}{TP/P \cdot P(Y = +) + (1 - TN/N) \cdot (1 - P(Y = +))}$$

$$\approx^* \frac{TP/P \cdot P/(P + N)}{TP/P \cdot P/(P + N) + (1 - TN/N) \cdot (1 - P/(P + N))} = \frac{TP}{TP + FP}$$

($*$) if $P(Y = +) \approx P/(P + N)$, i.e., if fraction of positives in the test set is same as population
Dataset selection

• Let $\Omega' = \Omega \times \{0, 1\}$, where:
  ▶ $(\omega, 1) \in \Omega'$ iff $\omega$ is selectionable in the dataset
  ▶ let $S$ be the name of the new feature

• Class independent selection:

\[
P(S = 1) = P(S = 1 | Y = +) = P(S = 1 | Y = -)
\]

• Class dependent selection

  ▶ Under-sampling negatives: $P(S = 1 | Y = -) < P(S = 1 | Y = +) = P(S = 1)$
  ▶ Over-sampling positives: $P(S = 1 | Y = +) > P(S = 1 | Y = -) = P(S = 1)$
  ▶ Data/distribution shift: $P(S = 1 | Y = -) \neq P(S = 1 | Y = +) \neq P(S = 1)$

• How confident are we that selection of our (training/test) dataset is class independent?
  ▶ Bias in data collection
  ▶ Change of distribution over time/domain

Then, confusion matrix is unpredictive of true precision/accuracy!
Precision of classifiers: correction under shift

<table>
<thead>
<tr>
<th>Predicted $\hat{Y}$</th>
<th>True $Y$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>$TP$</td>
<td>$PP$</td>
</tr>
<tr>
<td>−</td>
<td>$FN$</td>
<td>$PN$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>$P$</strong></td>
<td><strong>$N$</strong></td>
</tr>
<tr>
<td></td>
<td><strong>$P + N$</strong></td>
<td></td>
</tr>
</tbody>
</table>

When class dependent selection can occur?

- Undersampling $P(Y = +) \approx P/(P + \beta N)$ with $\beta = N_{\text{orig}}/N \geq 1$ rate in original dataset
- Oversampling $P(Y = +) \approx \alpha P/(\alpha P + N) = P/(P + N/\alpha)$ with $\alpha = P_{\text{orig}}/P \leq 1$
- Shift $P(Y = +) \approx \alpha P/((\alpha P + \beta N) = P/(P + \gamma N)$ with $\gamma = \beta/\alpha = (N_{\text{orig}}/P_{\text{orig}})/(N/P)$

What is the probability I really am positive given that I was predicted positive?  

$$P(Y = +|\hat{Y} = +) \approx \frac{TP/P \cdot P/(P + \gamma N)}{TP/P \cdot P/(P + \gamma N) + (1 - TN/N) \cdot (1 - P/(P + \gamma N))} = \frac{TP}{TP + \gamma FP}$$

Called $Prec = TP/(TP + FP)$, we have:

$$P(Y = +|\hat{Y} = +) \approx \frac{Prec}{Prec + \gamma(1 - Prec)}$$

Example: for $\gamma = 5$, $Prec = 0.9$, we have $P(Y = +|\hat{Y} = +) \approx 0.9/(0.9 + 5 \cdot 0.1) \approx 0.642$
### Accuracy of Classifiers

Consider the following confusion matrix:

<table>
<thead>
<tr>
<th></th>
<th>True Y</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Y</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>+</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>−</td>
<td>FN</td>
<td>TN</td>
</tr>
<tr>
<td>Total</td>
<td>P</td>
<td>N</td>
</tr>
</tbody>
</table>

- $P(\hat{Y} = +|Y = +) \approx \frac{TP}{P}$  
  
  [Sensitivity/Recall/TPR]
- $P(\hat{Y} = −|Y = −) \approx \frac{TN}{N}$  
  
  [Specificity/TNR]

**What is the probability that prediction is correct?**

$$P(\hat{Y} = Y) = P(\hat{Y} = +|Y = +)P(Y = +) + P(\hat{Y} = −|Y = −)P(Y = −) \approx (*)$$

$$\approx (*) \frac{TP}{P} \frac{P}{P + N} + \frac{TN}{N} \frac{N}{P + N} = \frac{TP + TN}{P + N}$$

(*) if $P(Y = +) \approx \frac{P}{P + N}$, i.e., if dataset selection is **class independent**!
Accuracy of classifiers: correction under shift

<table>
<thead>
<tr>
<th>Predicted $\hat{Y}$</th>
<th>True $Y$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>+</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>−</td>
<td>FN</td>
<td>TN</td>
</tr>
<tr>
<td>Total</td>
<td>$P$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

- Shift $P(Y = +) \approx \frac{\alpha P}{\alpha P + \beta N} = \frac{P}{P + \gamma N}$ with $\gamma = \frac{\beta}{\alpha} = \frac{(N_{\text{orig}}/P_{\text{orig}})}{(N/P)}$

What is the probability that prediction is correct? [Accuracy]

$$P(\hat{Y} = Y) = P(\hat{Y} = + | Y = +)P(Y = +) + P(\hat{Y} = - | Y = -)P(Y = -) \approx$$

$$\approx \frac{TP}{P} \frac{P}{P + \gamma N} + \frac{TN}{N} \frac{\gamma N}{P + \gamma N} = \frac{TP + \gamma TN}{P + \gamma N}$$

**Example:** for $\gamma = 10$, $P = N = 1000$, $TP = 950$, $TN = 800$:

$$Acc = \frac{(TP + TN)}{(P + N)} = .875$$

$$P(\hat{Y} = Y) = \frac{(TP + \gamma TN)}{(P + \gamma N)} \approx .814$$
Probabilistic classifier predictions: correction under shift

A probabilistic classifier predicts the posterior probability $P(Y = +|G = g, A = a)$.

Assume a biased posterior probability $\hat{S}((g, a)) \approx P(Y = +|S = 1, G = g, A = a)$, due to data shift.

**How to compute unbiased prediction** $P(Y = +|G = g, A = a)$?

- Class dependent selection, but feature independent selection:
  
  $$P(S = 1) \neq P(S = 1|Y = +) = P(S = 1|Y = +, G = g, A = a)$$

- From Bayes rule applied to $P'(\cdot) = P(\cdot|G = g, A = a)$:
  
  $$P'(Y = +|S = 1) = \frac{P'(Y = +)}{P'(Y = +) + \frac{P'(S=1|Y=-)}{P'(S=1|Y=+)}(1 - P'(Y = +))}$$

- For shift: $\frac{P'(S=1|Y=-)}{P'(S=1|Y=+)} \approx \frac{N/N_{\text{orig}}}{P/P_{\text{orig}}} = \alpha/\beta = 1/\gamma$, hence:

  $$P'(Y = +|S = 1) = \frac{P'(Y = +)}{P'(Y = +) + (1 - P'(Y = +))/\gamma}$$
Probabilistic classifier predictions: correction under shift

A probabilistic classifier predicts the posterior probability \( P(Y = +|G = g, A = a) \)

Assume a biased posterior probability \( \hat{S}((g, a)) \approx P(Y = +|S = 1, G = g, A = a) = P'(Y = +|S = 1) \)

**How to compute unbiased prediction \( P(Y = +|G = g, A = a) \)?**

- and then, solving for \( P'(Y = +) = P(Y = +|G = g, A = a) \):
  \[
P'(Y = +) = \frac{P'(Y = +|S = 1)}{P'(Y = +|S = 1) + \gamma(1 - P'(Y = +|S = 1))}
\]

- Correction under shift:
  \[
  \frac{\hat{S}((g, a))}{\hat{S}((g, a)) + \gamma(1 - \hat{S}((g, a)))}
  \]

*Same formula as for precision!*
Optional references

Optional readings: [Pozzolo, 2015], [Sipka, 2021] (consider the case when $\gamma$ is unknown)

Andrea Dal Pozzolo, Olivier Caelen, and Gianluca Bontempi (2015)
When is Undersampling Effective in Unbalanced Classification Tasks?
ECML/PKDD (1) 200–215.
Lecture Notes in Computer Science, volume 9284.
https://doi.org/10.1007/978-3-319-23528-8_13

Tomas Sipka, Milan Sulc, and Jiri Matas (2021)
The Hitchhiker’s Guide to Prior-Shift Adaptation.
CoRR abs/2106.11695.
https://arxiv.org/abs/2106.11695