

Master Program in *Data Science and Business Informatics*

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.

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

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

- $P(+|C) = 0.99$ *[Sensitivity/Recall/True Positive Rate]*
- $P(-|C^c) = 0.99$ *[Specificity/True Negative Rate]*

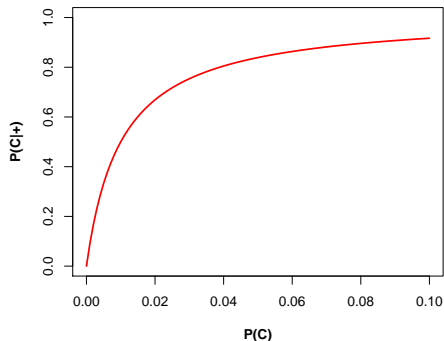
What is the probability I really have Covid-19 given that I tested positive? *[Precision]*

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

$$P(C|+) = \frac{0.99 \cdot P(C)}{0.99 \cdot P(C) + 0.01 \cdot (1 - P(C))}$$

Testing for Covid-19

$P(C)$, the probability of having Covid-19, is **unknown**. Let's plot $P(C|+)$ over $P(C)$:



- For $P(C) = 0.02$, $P(C|+) = .67$
- For $P(C) = 0.06$, $P(C|+) = .86$
- For $P(C) = 0.10$, $P(C|+) = .92$

See R script

Bayes' Rule

BAYES' RULE. Suppose the events C_1, C_2, \dots, C_m are disjoint and $C_1 \cup C_2 \cup \dots \cup C_m = \Omega$. The conditional probability of C_i , given an arbitrary event A , can be expressed as:

$$P(C_i | A) = \frac{P(A | C_i) \cdot P(C_i)}{P(A | C_1)P(C_1) + P(A | C_2)P(C_2) + \dots + P(A | C_m)P(C_m)}.$$

- 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 $j = 1, \dots, m$
 - ▶ E.g., in classification problems (see Bayesian classifiers from Data Mining)
- $P(C_i)$ is called the *prior* probability
- $P(C_i | A)$ is called the *posterior* probability (after seeing event A)

(Machine Learning) 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 $(Y = +)^c$
- 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 $(\hat{Y} = +)^c$
- $P(Y = \hat{Y})$, i.e., $P(Y = + \cap \hat{Y} = +) + P(Y = - \cap \hat{Y} = -)$ *[True Accuracy]*
- $P(Y = + \mid \hat{Y} = +)$ *[True Precision]*
- $P(\hat{Y} = + \mid Y = +)$ *[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!

		True Y		Total
		+	-	
Predicted \hat{Y}	+	TP	FP	PP
	-	FN	TN	PN
Total		P	N	P + N

- $P(\hat{Y} = + | Y = +) \approx TP/P$
- $P(\hat{Y} = - | Y = -) \approx TN/N$
- “ \approx ” reads as “approximately”

[Sensitivity/Recall/TPR]

[Specificity/TNR]

[Probability estimation]

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

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

Precision of classifiers

Confusion matrix over the test set!

		True Y		Total
		+	-	
Predicted \hat{Y}	+	TP	FP	PP
	-	FN	TN	PN
Total		P	N	P + N

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

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

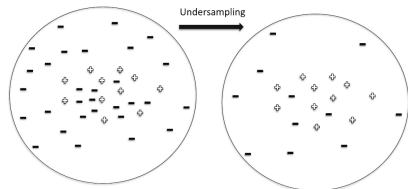
$$\begin{aligned} P(Y = + | \hat{Y} = +) &= \frac{P(\hat{Y} = + | Y = +) \cdot P(Y = +)}{P(\hat{Y} = + | Y = +) \cdot P(Y = +) + (1 - P(\hat{Y} = - | Y = -)) \cdot P(Y = -)} \\ &\approx \frac{TP/P \cdot P(Y = +)}{TP/P \cdot P(Y = +) + (1 - TN/N) \cdot P(Y = -)} \\ &\stackrel{(*)}{\approx} \frac{TP/P \cdot P/(P + N)}{TP/P \cdot P/(P + N) + (1 - TN/N) \cdot N/(P + N)} = \frac{TP}{TP + FP} \end{aligned}$$

(*) 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 ω is selected 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: $P(S = 1|Y = -) \neq P(S = 1|Y = +) \neq P(Y = +)$
- How confident are we that selection of our (training/test) dataset is class independent?
 - ▶ Bias in data collection
 - ▶ Change of distribution over time/domain

[Selection bias]
[Distribution shift]

Then, confusion matrix is uninformative of true precision/accuracy!

Precision of classifiers: correction under shift

		True Y		Total
		+	-	
Predicted \hat{Y}	+	TP	FP	PP
	-	FN	TN	PN
Total		P	N	P + N

When class dependent selection can occur?

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

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

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

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

		True Y		
		+	-	Total
Predicted \hat{Y}	+	<i>TP</i>	<i>FP</i>	<i>PP</i>
	-	<i>FN</i>	<i>TN</i>	<i>PN</i>
	Total	<i>P</i>	<i>N</i>	<i>P + N</i>

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

[Sensitivity/Recall/TPR]

- $P(\hat{Y} = - | Y = -) \approx TN/N$

[Specificity/TNR]

What is the probability that prediction is correct?

[Accuracy]

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

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

(*) if $P(Y = +) \approx P/(P+N)$, i.e., if dataset selection is **class independent!**

Accuracy of classifiers: correction under shift

		True Y		
		+	-	Total
Predicted \hat{Y}	+	<i>TP</i>	<i>FP</i>	<i>PP</i>
	-	<i>FN</i>	<i>TN</i>	<i>PN</i>
	Total	<i>P</i>	<i>N</i>	<i>P + N</i>

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

What is the probability that prediction is correct?

[Accuracy]

$$\begin{aligned} P(\hat{Y} = Y) &= P(\hat{Y} = + | Y = +)P(Y = +) + P(\hat{Y} = - | Y = -)P(Y = -) \approx \\ &\approx \frac{TP}{P} \frac{P}{P + \gamma N} + \frac{TN}{N} \frac{\gamma N}{P + \gamma N} = \frac{TP + \gamma TN}{P + \gamma N} \end{aligned}$$

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

$$Acc = (TP + TN) / (P + N) = .875$$

$$P(\hat{Y} = Y) = (TP + \gamma TN) / (P + \gamma N) \approx .814$$

Probabilistic classifier predictions: correction under shift

Assume a *biased* posterior probability $P(Y = +|S = 1, X = x)$, due to data shift, returned by a probabilistic classifier *[predict_proba in Python]*

How to compute unbiased prediction $P(Y = +|X = x)$?

- Class dependent selection, but feature independent selection:

$$P(S = 1) \neq P(S = 1|Y = +) = P(S = 1|Y = +, X = x)$$

- From Bayes rule applied to $P'(\cdot) = P(\cdot|X = x)$:

$$P'(Y = +|S = 1) = \frac{P'(Y = +)}{P'(Y = +) + \frac{P'(S=1|Y=-)}{P'(S=1|Y=+)}(1 - P'(Y = +))}$$

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



$$P'(Y = +|S = 1) = \frac{P'(Y = +)}{P'(Y = +) + (1 - P'(Y = +))/\gamma}$$

and then: $P(Y = +|X = x) = P'(Y = +) = \frac{P'(Y=+|S=1)}{P'(Y=+|S=1) + \gamma(1 - P'(Y=+|S=1))}$

Same formula as for precision!

Optional references

Optional readings: [Pozzolo, 2015], [Sipka, 2021] (consider the case when γ 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>