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 tosses
- $A = \{\text{first coin is H}\} = \{(H, H), (H, T)\} \quad P(A) = 1/2$
- $B = \{\text{second coin is H}\} = \{(H, H), (T, H)\} \quad P(B) = 1/2$

$$P(A \cap B) = 1/4 = P(A)P(B)$$

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

$$P(A \cap B|C) = \frac{P(A \cap B \cap C)}{P(C)} = 1/2 \neq P(A|C)P(B|C) = \frac{P(A \cap C)}{P(C)} \cdot \frac{P(B \cap C)}{P(C)} = 1/4$$

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

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:

$$P(A|C)P(B|C) = \frac{P(A \cap C)P(B \cap C)}{P(C)^2} = \frac{P(A)P(C)P(B)P(C)}{P(C)^2} = P(A)P(B)$$

An application to machine learning classifiers

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

Can be rewritten as if $P(A|B) = P(A)$ and $P(A|B \cap C) \neq P(A|C)$

- $\Omega = \{\text{summer, winter}\} \times \{\text{long-hair, short-hair}\} \times \{\text{eat-icecream, dont-eat-icecream}\}$
- $A = \{(\_, \_, \text{eat-icecream})\}$
- $B = \{(\_, \text{long-hair}, \_)\}$
- $C = \{(\text{summer}, \_, \_)\}$

How do we read the result above?

- if $P(A|B) = P(A)$ read as "*long-hair is not predictive of eating ice cream*"
- if $P(A|B \cap C) \neq P(A|C)$ read as "*in the summer, long-hair is predictive of eating ice cream*"

What can we conclude in general for features of machine learning classifiers?

- A feature can be non-relevant in isolation, but relevant together other featurs
- We cannot do feature selection by looking at a single feature at a time!

Testing for Covid-19

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

- $\Omega = \{$ people aged 18 or higher $\}$
- $+ = \{$ people tested positive $\}$ $\quad - = \{$ people tested negative $\} = +^c$
- $C = \{$ people with Covid-19 $\}$ $\quad C^c = \{$ 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]*

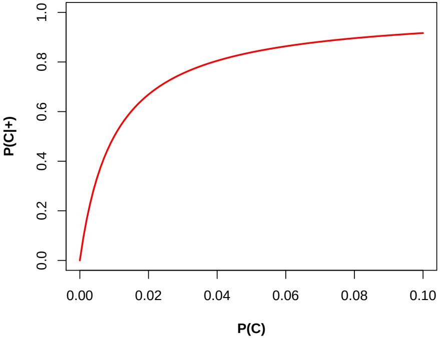
$$P(C|+) = \frac{P(C \cap +)}{P(+)} = \frac{P(+|C) \cdot P(C)}{P(+)} = \frac{P(+|C) \cdot P(C)}{P(+|C) \cdot P(C) + P(+|C^c) \cdot P(C^c)}$$

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

$P(C)$ is unknown!

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 $C_1, C_2, \ldots, C_m$ are disjoint and $C_1 \cup C_2 \cup \cdots \cup C_m = \Omega$. The conditional probability of C_i, given an arbitrary event A, can be expressed as:

$$\mathrm{P}(C_i \mid A) = \frac{\mathrm{P}(A \mid C_i) \cdot \mathrm{P}(C_i)}{\mathrm{P}(A \mid C_1)\mathrm{P}(C_1) + \mathrm{P}(A \mid C_2)\mathrm{P}(C_2) + \cdots + \mathrm{P}(A \mid C_m)\mathrm{P}(C_m)}.$$

- 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 $j = 1, \ldots, m$
  - 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

- $\Omega = \{\text{f, m}\} \times \mathbb{N} \times \{+, -\}$
- 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: $\hat{Y} : \{\text{f, m}\} \times \mathbb{N} \rightarrow \{+, -\}$ 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!

		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]*

$$P(Y=+|\hat{Y}=+) = \frac{TP}{TP+FP}$$ is it always this?

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]*

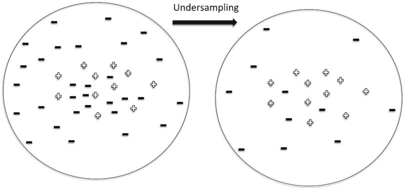
$$P(Y=+|\hat{Y}=+) = \frac{P(\hat{Y}=+|Y=+)\cdot P(Y=+)}{P(\hat{Y}=+|Y=+)\cdot P(Y=+) + (1-P(\hat{Y}=-|Y=-))\cdot P(Y=-)}$$

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

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

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

Dataset selection



- Let $\Omega = \{\text{f, m}\} \times \mathbb{N} \times \{+, -\} \times \{0, 1\}$, where:
  - $S = v$ is $\{\omega \in \Omega \mid \omega = (\_, \_, \_, v)\}$
  - selected ($S = 1$) or not ($S = 0$) in the observed dataset
- Typical assumption: class independent selection:

$$P(S = 1) = P(S = 1|Y = +) = P(S = 1|Y = -)$$

- Reasons for class dependent selection:
  - Bias in data collection *[Selection bias]*
  - Change of distribution over time/domain *[Distribution shift]*

  Confusion matrix (over test set) is unpredictive of true precision/accuracy (over the population)!
- Forms of class dependent selection
  - Under-sampling negatives: $P(S = 1|Y = -) < P(S = 1|Y = +) = P(S = 1)$
  - Over-sampling positives: $P(S = 1|Y = +) > P(S = 1|Y = -) = P(S = 1)$
  - Prior probability shift: $P(S = 1|Y = -) \neq P(S = 1|Y = +) \neq P(S = 1)$

Dataset selection

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

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

What is P(Y=+)? By the Bayes' rule:

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

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

where $\gamma = \frac{P(S=1|Y=+)}{P(S=1|Y=-)}$ By solving back w.r.t. $P(Y=+)$, we have:

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

since $P(Y=+|S=1) \approx P/(P+N)$.

Odds ratio γ

By the Bayes' rule (2 times):

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

is called the **odds ratio**.

- odds in the sample space (population): $\frac{P(Y=-)}{P(Y=+)} \approx \frac{N_{pop}}{P_{pop}}$ is unknown
- odds in the selected: $\frac{P(Y=-|S=1)}{P(Y=+|S=1)} \approx \frac{N}{P}$

In general, $\gamma \approx (N_{pop}/P_{pop})/(N/P)$ with N_{pop} and P_{pop} from an unbiased dataset.

- Prior shift $P(Y=+) \approx P/(P+\gamma N)$
- Undersampling $P(Y=+) \approx P/(P+\beta N)$ with $\beta = N_{pop}/N \geq 1$
- Oversampling $P(Y=+) \approx P/(P+N/\alpha)$ with $\alpha = P_{pop}/P \leq 1$

We do not know γ but:

- we can estimate/approximate it
- we can reason hypothetically on possible range of values for it.

Precision of classifiers: correction under shift

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

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 also have (*correction of precision*):

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

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

See R script

Accuracy of classifiers

		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]*

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

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

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

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

Accuracy of classifiers: correction under shift

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

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

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

$Acc = (TP+TN)/(P+N) = .875$ $\qquad$ $P(\hat{Y}=Y) = (TP+\gamma TN)/(P+\gamma N) \approx .814$

Probabilistic classifier predictions: correction under shift

A probabilistic classifier intended to predict the posterior probability $P(Y=+|G=g, A=a)$
[predict_proba in Python]

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 $P'(\cdot) = P(\cdot|G=g, A=a)$, and following the same reasoning as for precision, correction under prior probability shift is:

$$P(Y=+|G=g, A=a) = \frac{\hat{S}((g,a))}{\hat{S}((g,a)) + \gamma(1-\hat{S}((g,a)))}$$

- *Same formula as for precision!*

Optional references

Optional readings:

- [Sipka et al., 2022] survey methods for prior-shift adaptation (also when γ is unknown!).
- [Pozzolo et al., 2015] apply correction to the study of effectiveness of undersampling.

,

Tomáš Šipka, Milan Šulc, and Jiří Matas (2022)
The Hitchhiker's Guide to Prior-Shift Adaptation.
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 1516-1524.
https://arxiv.org/abs/2106.11695

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