Weekly Overview Slides of Statistical Machine Learning CSE 575, Spring 2023

Moses A. Boudourides[1]

SPA and SCAI
Arizona State University

[1] Moses.Boudourides@asu.edu

Week 7

Examples on Classification via the Statistical Approach

February 23, 2023

Bayes Theorem

Bayes Theorem

- Let h be a hypothesis and $p(h)$ be the **prior probability** that h holds.
- Let $\mathcal{D}$ be the training data and $p(\mathcal{D})$ be the **prior probability** that $\mathcal{D}$ will be observed (i.e., the probability of $\mathcal{D}$ given no knowledge which hypothesis holds).
- Let $p(\mathcal{D}|h)$ be the probability of observing data $\mathcal{D}$ given some world in which hypothesis h holds. $p(\mathcal{D}|h)$ is called the **likelihood** of h for $\mathcal{D}$.
- Let $p(h|\mathcal{D})$ be the probability that h holds given the observed training data $\mathcal{D}$. $p(h|\mathcal{D})$ is called the **posterior probability** of h, because it reflects our confidence that h holds after we have seen the training data $\mathcal{D}$.
- **Bayes Theorem**:

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{p(\mathcal{D})}.$$

Bayes Classification

MAP Bayes Classifier

Given data $\mathcal{D} = \boldsymbol{x}$ (which is a vector of features), a set of K classes $C_j, j = 1, \ldots, K$, and the respective *posterior probabilities* $p(C_j|\boldsymbol{x})$, classify $\boldsymbol{x}$ by assigning it to the class C_k, for which the posterior probability becomes maximum (thus, C_k is called class with *maximum a posteriori probability* or *MAP*):

$$C_k = \underset{j=1,\ldots,K}{\arg\max}\, p(C_j|\boldsymbol{x}),$$

or, employing Bayes Theorem,

$$C_k = \underset{j=1,\ldots,K}{\arg\max}\, p(\boldsymbol{x}|C_j)p(C_j),$$

where $p(\boldsymbol{x}|C_j)$ denotes the class likelihood and $p(C_j)$ the class prior.

The **total error probability of a Bayes classifier** is

$$E_{\mathrm{Bayes}} = \int p(\boldsymbol{x}) \min_{j=1,\ldots,K}\{p(C_j|\boldsymbol{x})\}\, d\boldsymbol{x} \;\leq \min_{j=1,\ldots,K}\{p(C_j)\},$$

where $\{p(C_j)\}$ are the *a priori probabilities* of the classes.

Example

In a medical diagnosis situation, there two cases, either *cancer* or $\neg$*cancer*, and the lab data need to be classified as either positive ($\oplus$) or negative ($\ominus$). We know $p(cancer) = 0.008$ and $p(\neg cancer) = 0.992$. Moreover, for the lab test, we know that $p(\oplus|cancer) = 0.98 \Rightarrow p(\ominus|cancer) = 0.02$ and that $p(\oplus|\neg cancer) = 0.03 \Rightarrow p(\ominus|\neg cancer) = 0.97$.
(a) If a new case is observed with a positive test result, should it be classified as having cancer or not?

Solution

The maximum a posteriori (MAP) hypothesis after the lab test yields

$$p(\oplus|cancer) \cdot p(cancer) = 0.98 \times 0.008 = 0.0078,$$
$$p(\oplus|\neg cancer) \cdot p(\neg cancer) = 0.03 \times 0.992 = 0.0298.$$

Thus, since the MAP hypothesis is

$$h_{\mathrm{MAP}} = \arg\max_{h \in \{cancer, \neg cancer\}} p(\mathcal{D}|h) \cdot p(h),$$

we get for the new case that $h_{\mathrm{MAP}} = \neg cancer$. In particular, after noramization, we have

$$p(cancer|\oplus) = \frac{0.0078}{0.0078 + 0.0298} = 0.21,$$
$$p(\neg cancer|\oplus) = \frac{0.0298}{0.0298 + 0.0078} = 0.79.$$

Example (cont.)

(b) Knowing that the previous lab test was imperfect, a second test (assumed to be **independent** of the former) is conducted. If the second test has again returned a positive result, should it be classified as having cancer or not?

Solution

Because of independence, the maximum a posteriori (MAP) hypothesis after the second test yields

$$\begin{aligned} p(\oplus\oplus|cancer)\cdot p(cancer) &= p(\oplus|cancer)\cdot p(\oplus|cancer)\cdot p(cancer)\\ &= 0.98\times 0.98\times 0.008 = 0.007644,\\ p(\oplus\oplus|\neg cancer)\cdot p(\neg cancer) &= p(\oplus|\neg cancer)\cdot p(\oplus|\neg cancer)\cdot p(\neg cancer)\\ &= 0.03\times 0.03\times 0.992 = 0.000894. \end{aligned}$$

Thus, we get for the second case that $h_{\mathrm{MAP}} = cancer$. In particular, after noramization, we have

$$p(cancer|\oplus\oplus) = \frac{0.007644}{0.007644+0.000894} = 0.895,$$

$$p(\neg cancer|\oplus\oplus) = \frac{0.000894}{0.000894+0.007644} = 0.105.$$

Exercise

In some elections, $1,000,000$ people are voting for either candidate A or candidate B. However, there are $1,000$ people who have already voted by postal voting and all of them have voted for candidate A. Assuming that all the remaining voters are voting by flipping a (non–manipulated) coin, what is the probability that candidate A wins the elections?

Is it 0.5005?

Discrete Bayes Classification

Example

Let a dataset of pencils, either yellow (Y) or white (W), to be classified in two categories: pencils (C_1) or graphite (C_2). If $p(C_1) = 1/3, p(C_2) = 2/3, p(Y|C_1) = 1/5, p(W|C_1) = 4/5, p(Y|C_2) = 2/3, p(W|C_2) = 1/3$, compare the total error probability of the Bayes classifier with the a priori probabilities of the two categories of pencils. **What is the risk of that decision?**

Solution

$$p(Y) = p(C_1)p(Y|C_1) + p(C_2)p(Y|C_2) = \frac{1}{3}\cdot\frac{1}{5} + \frac{2}{3}\cdot\frac{2}{3} = \frac{23}{45},$$

$$p(W) = p(C_1)p(W|C_1) + p(C_2)p(W|C_2) = \frac{1}{3}\cdot\frac{4}{5} + \frac{2}{3}\cdot\frac{1}{3} = \frac{22}{45},$$

implying that $p(C_1|Y) = \frac{p(C_1)p(Y|C_1)}{p(Y)} = \frac{(1/3)\cdot(1/5)}{(23/45)} = \frac{3}{23}$, $p(C_1|W) = \frac{p(C_1)p(W|C_1)}{p(W)} = \frac{(1/3)\cdot(4/5)}{(22/45)} = \frac{6}{11}$, $p(C_2|Y) = 1 - p(C_1|Y) = \frac{20}{23}$ and $p(C_2|W) = 1 - p(C_1|W) = \frac{5}{11}$.Thus, the total error is

$$p(Y)\cdot\min_{i=1,2}\{p(C_i|Y)\} + p(W)\cdot\min_{i=1,2}\{p(C_i|W)\} = \frac{23}{45}\cdot\frac{3}{23} + \frac{22}{45}\cdot\frac{5}{11} = \frac{13}{45}$$

$$< \frac{1}{3} = \min_{i=1,2}\{p(C_i)\}.$$

Continuous Bayes Classification

Example

Consider the previous example, but instead of two colors, assume all pencils to be yellow of shade varying from 0 to 2 w.r.t. the following conditional probability distributions of the shade x:

$$p(x|C_1) = -\frac{x}{2} + 1 \text{ and } p(x|C_2) = \frac{x}{2}, \text{ for } x \in [0, 2].$$

Thus, the a priori probability distribution of x is

$$\begin{aligned} p(x) = p(C_1)p(x|C_1) + p(C_2)p(x|C_2) &= \frac{1}{3}(-\frac{x}{2} + 1) + \frac{2}{3}\frac{x}{2} \\ &= \frac{1}{6}(x + 2). \end{aligned}$$

Therefore, the a posteriori probabilities are

$$p(C_1|x) = \frac{\frac{1}{3} \cdot (1 - \frac{x}{2})}{\frac{1}{6}(x+2)} = \frac{2-x}{2+x},$$

$$p(C_2|x) = \frac{\frac{2}{3} \cdot \frac{x}{2}}{\frac{1}{6}(x+2)} = \frac{2x}{2+x}.$$

Now, the total error probability is found to be (as expected):

$$E_{\mathrm{Bayes}} = \int_0^2 \frac{1}{6}(x+2) \min\{\frac{1}{3}, \frac{2}{3}\}\, dx = \frac{1}{3} \leq \min\{\frac{1}{3}, \frac{2}{3}\}.$$

What is the risk of that decision? What are the corresponding decision regions?

Naive Bayes Classification

Naive Bayes Classifier

Given a set of K classes $C_j, j = 1, \ldots, K$, and data being an D–dimensional vector of features $\boldsymbol{x} = (x_1, \ldots, x_D)$, if the components (features) of $\boldsymbol{x}$ are *statistically independent* and, hence, the joint PDF of the likelihood is written as a product of D marginals,

$$p(\boldsymbol{x}|C_j) = \prod_{i=1}^{D} p(x_i|C_j), j = 1, \ldots, K,$$

then the **naive Bayes classifier** assigns to $\boldsymbol{x}$ the class C_k such that

$$C_k = \underset{j=1,\ldots,K}{\arg\max}\, p(C_j) \prod_{i=1}^{D} p(x_i|C_j).$$

Example

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*Play Tennis?*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example (cont.)

Training data:

- $n = 14$ cases (days) of training data, with:
- $D = 4$ features $\{Outlook, Temperature, Humidity, Wind\}$, to be classified in:
- $K = 2$ labels (classes) $\{Play\ \ Tennis = yes, Play\ \ Tennis = no\}$.

Target: Classify (whether to play tennis or not) the following novel instance:

$(Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)$.

A priori target probabilities:

$$p(Play\ \ Tennis = yes) = \frac{9}{14} = 0.64,$$
$$p(Play\ \ Tennis = no) = \frac{5}{14} = 0.36.$$

Example (cont.)

Conditional probabilities:

$$p(Outlook = sunny | Play\ \ Tennis = yes) = \frac{2}{5} = 0.4,$$

$$p(Outlook = sunny | Play\ \ Tennis = no) = \frac{3}{5} = 0.6,$$

$$p(Temperature = cool | Play\ \ Tennis = yes) = \frac{3}{4} = 0.75,$$

$$p(Temperature = cool | Play\ \ Tennis = no) = \frac{1}{4} = 0.25,$$

$$p(Humidity = high | Play\ \ Tennis = yes) = \frac{3}{7} = 0.43,$$

$$p(Humidity = high | Play\ \ Tennis = no) = \frac{4}{7} = 0.57,$$

$$p(Wind = strong | Play\ \ Tennis = yes) = \frac{3}{6} = 0.5,$$

$$p(Wind = strong | Play\ \ Tennis = no) = \frac{3}{6} = 0.5.$$

Estimates of Naive Bayes Classifier: (omitting attribute names for brevity)

$$p(yes)\ p(sunny|yes)\ p(cool|yes)\ p(high|yes)\ p(strong|yes) = 0.64 \times 0.4 \times 0.75 \times 0.43 \times 0.5 = 0.04128,$$

$$p(no)\ p(sunny|no)\ p(cool|no)\ p(high|no)\ p(strong|no) = 0.36 \times 0.6 \times 0.25 \times 0.57 \times 0.5 = 0.01539.$$

Conclusion: The Naive Bayes Classifier assigns the target value $\{Play\ \ Tennis = yes\}$ to this new instance, based on the probability estimates learned from the training data. Furthermore, by normalizing the above values, the probability of the estimated target value *yes* is $\frac{0.04128}{0.01539+0.04128} = 0.728$.