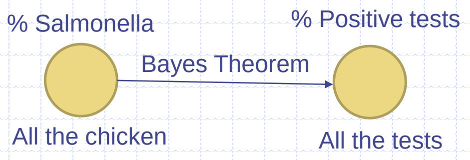
INM431 Machine Learning

Artur S. d'Avila Garcez
a.garcez@city.ac.uk
http://www.staff.city.ac.uk/~aag/

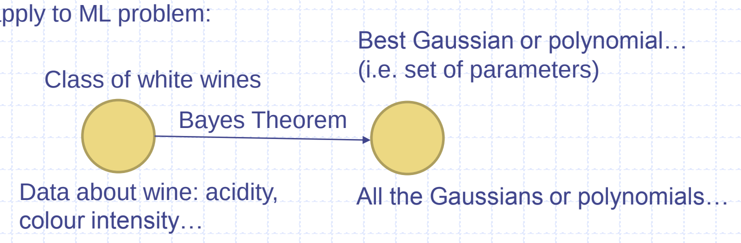
Content

- Naïve Bayes classifiers

ML problem (in a nutshell)



Now, apply to ML problem:



Wine dataset: https://archive.ics.uci.edu/ml/datasets/Wine

Naïve Bayes

- A popular baseline method for text classification with assumption of independence among variables
- Given $\mathbf{x} = (x_1, \ldots, x_n)$ representing n variables (features), calculating the probability tables is intractable with large n (e.g. words appearing in a document), where k below is the number of document classes/types

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}.$$

- Under **maximum-likelihood** this can be done by evaluating an expression in linear time, rather than by iterative approximation…
- **Scalable**, requiring a number of parameters linear on the number of variables (e.g. word frequencies)

Naïve Bayes (cont.)

With the Naïve conditional independence assumption:

$$\begin{aligned} p(C_k|x_1,\dots,x_n) &\propto p(C_k, x_1, \dots, x_n) \\ &\propto p(C_k)\ p(x_1|C_k)\ p(x_2|C_k)\ p(x_3|C_k)\ \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i|C_k)\,. \end{aligned}$$

$$p(C_k|x_1,\dots,x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \quad \text{where} \quad Z = p(\mathbf{x})$$

Naïve Bayes classifier

Combines Naïve Bayes model with a **decision rule**, e.g. *maximum a posteriori* or MAP decision rule, which selects the most probable hypothesis:

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

Where did Z go? The partition function Z can be removed since results won't need to be normalized before the decision rule (argmax) is applied

Example (1)

chills	runny nose	headache	fever	Flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

chills	runny nose	headache	fever	Flu?
Y	N	Mild	N	?

https://www.youtube.com/watch?v=ZAfarappAO0

Example (2) Flu=Y/N?

Prior:

P(flu) = 5/8, P(~flu) = 3/8

Likelihoods:

P(chills|flu) = 3/5 P(~chills|flu) = 2/5

P(runny|flu) = 4/5 P(~runny|flu) = 1/5

P(mild|flu) = 2/5 P(no|flu) = 1/5 P(strong|flu) = 2/5

P(fever|flu) = 4/5 P(~fever|flu) = 1/5

Posterior (1):

P(flu|chills,~runny,mild,~fever) =
P(flu)P(chills|flu)P(~runny|flu)P(mild|flu)P(~fever|flu) =
0.625 x 0.6 x 0.2 x 0.4 x 0.2 = 0.006

Example (3) Flu=Y/N?

Prior:

P(flu) = 5/8,   P(~flu) = 3/8

More Likelihoods:

P(chills|~flu) = 1/3   P(~chills|~flu) = 2/3

P(runny|~flu) = 1/3 P(~runny|~flu) = 2/3

P(mild|~flu) = 1/3   P(no|~flu) = 1/3   P(strong|~flu) = 1/3

P(fever|~flu) = 1/3   P(~fever|~flu) = 2/3

Posterior (2):

P(~flu|chills,~runny,mild,~fever) = 3/8 x 1/3 x 2/3 x 1/3 x 2/3

Prediction:

argmax (P(flu),P(~flu)) = argmax (0.006, 0.0185) = No Flu!

Try this for other test examples…

Naïve Bayes family of classifiers

A **class prior** may be calculated by assuming equiprobable classes: prior = 1 / (number of classes), or by calculating an estimate for the class probability from the training set:
class prior = (number of samples in the class) / (total number of samples)

Variations: Gaussian, multinomial, Bernoulli naïve Bayes, etc.

Despite the naïve conditional independence assumption, naïve Bayes classifiers can be surprisingly efficient on various datasets…

Gaussian naïve Bayes

Priors are calculated as before…

Likelihoods can be calculated from the training set by finding mean and variance for each attribute given a class

Posterior is calculated as before but using the following equation in the case of continuous variable x taking value v:

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

Regularization

What if just one of many conditional probabilities in

$$\hat{y} = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}}\ p(C_k)\prod_{i=1}^{n} p(x_i|C_k).$$

is equal to zero?

Use Laplace (a.k.a. "add 1") smoothing:

Let **x** = (x_1,…,x_d) be observation from a multinomial distribution with N trials (x_i is the number of times outcome i is observed)

A smoothed version of each x_i is given by $(x_i+1)/(N+d)$

The resulting estimate will be between the empirical probability (relative frequency) x_i / N and the uniform probability 1/d

Continuous and discrete data

- Since we have the conditional independence assumption in Naive Bayes, mixing variables is not a problem.
- We can compute the likelihoods of binary variables using a Bernoulli distribution, and compute the likelihoods of the continuous variables with a Gaussian.