Introduction to Machine Learning
Naïve Bayes

Nikolay Manchev

1 June 2017

London Machine Learning Study Group

Housekeeping

Thanks to our sponsor

www.opentable.com

Next events

http://www.meetup.com/London-Machine-Learning-Study-Group

Follow me

https://twitter.com/nikolaymanchev

Slides and code

Available at https://github.com/nmanchev/MachineLearningStudyGroup

Previous recordings

Available at https://www.youtube.com/c/NikolayManchev

Machine Learning Models

So far we looked at

- Linear Regression
- Polynomial Regression
- Decision Trees
- Logistic Regression

Topic for today

- Naïve Bayes

Definition

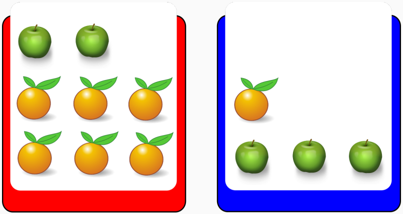
Rules of Probability

- Sum rule

$$P(X) = \sum_Y p(X, Y)$$

- Product rule

$$P(X, Y) = P(Y|X)P(X)$$



Simple example given by [Bis06]

Given that $P(B = r) = \frac{4}{10}$ and $P(B = r) = \frac{6}{10}$, we use the rules to solve problems like $P(F = a) =?$ and $P(B = r, F = a) =?$

$P(F = a) = \sum_{B \in \{r,b\}} P(F = a|B)$
$P(B = r, F = a) = P(F = a|B = r) \times P(B = r)$

Bayes Rule

Starting with the product rule

$$P(X, Y) = P(Y|X)P(X) \tag{1}$$

We can swap X and Y and rewrite it as

$$P(Y, X) = P(X|Y)p(Y) \tag{2}$$

We can solve (1) for $P(Y|X)$

$$P(Y|X) = \frac{P(X, Y)}{P(X)} \tag{3}$$

The symmetry rule tells us that $P(X, Y) = P(Y, X)$ so we can replace $P(X, Y)$ with (2) to get

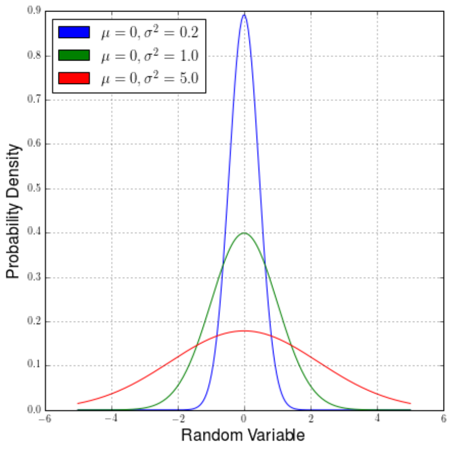
$$P(Y|Y) = \frac{P(X|Y)P(Y)}{P(X)} \leftarrow \text{This is Bayes' theorem} \tag{4}$$

Gaussian Distribution

- A very common continuous probability distribution

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

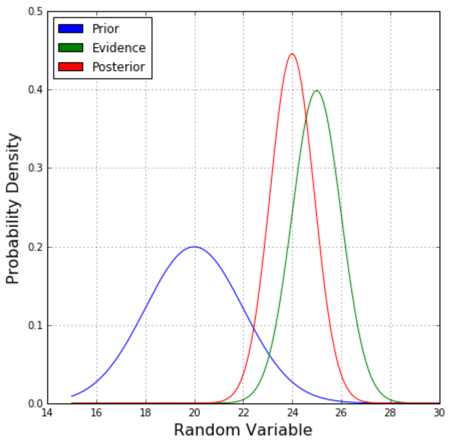
- Occurs naturally in many situations
    - height of people
    - salaries
    - blood pressure
- CLT: Good approximation for the sum or the means of many processes



Bayesian Inference

A method of inference where the probability of a hypothesis is updated as new evidence becomes available.

- Begin with a **prior** distribution processes
- Collect data (**E**) to obtain the observed distribution
- Calculate the **likelihood** – how compatible is is the evidence with the hypothesis
- Obtain the **posterior** – the probability of our hypothesis given the observed evidence



Computing the posterior

Bayes' Theorem revisited

We can use Bayes' theorem to express $P(H|E)$

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

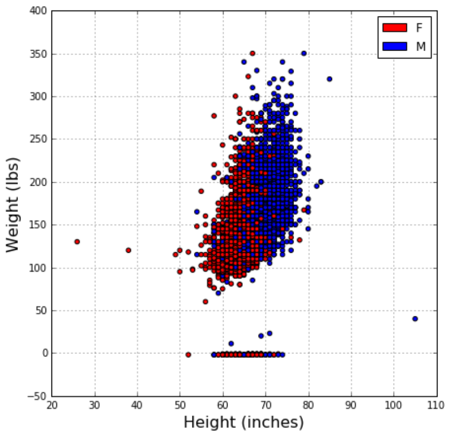
where

- $P(H|E)$ is our **posterior**
- $P(E|H)$ is the **likelihood**
- $P(H)$ is the **prior**
- $P(E)$ is a normalizing constant (reduces the probability function to a probability density function with total probability of 1)

Example [1/3]

Data from National Longitudinal Youth Survey, Bureau of Labor Statistics, United States Department of Labor [oL96]

Sex	Height	Weight
1	67	150
0	67	140
1	67	100
1	62	185
0	69	145
1	68	140
...	...	...



Example[2/3]

Sex	Height	Weight
0	67	140
0	69	145
0	69	183
0	71	175
0	66	108

Sex	Height	Weight
1	67	150
1	67	100
1	62	185
1	68	140
1	64	123

We have two possible classes $C \in \{M, F\}$. The priors are $P(C = M) = \frac{10}{5} = 0.5$ and $P(C = F) = 0.5$. We then compute the statistics for each subclass:

$\mu_M = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{67+69+\cdots+66}{5} = 68.4$, $\mu_F = 65.60$

$\sigma_M = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu_M)^2}{n}} = 1.74$, $\sigma_F = 2.24$

Example[3/3]

We can now compute the likelihood (i.e. how likely it is that an observation came from class M or F)

Say we have a measure of 69 inches for the **height** attribute.

The likelihood this measure came from class M is then

$$P(x = 69|C = M) = \frac{1}{\sqrt{2\pi\sigma_M^2}} e^{-\frac{(69-\mu_M)^2}{2\sigma_M^2}}$$

Respectively, the likelihood this measure came from class F is

$$P(x = 69|C = F) = \frac{1}{\sqrt{2\pi\sigma_F^2}} e^{-\frac{(69-\mu_F)^2}{2\sigma_F^2}}$$

Having the prior and likelihood allows us to obtain the posterior.

MAP

Maximum a posteriori estimation

Our prediction is the value of C, which maximizes the posterior distribution.

$$C_{MAP} = \arg\max_{c \in C} \frac{P(x|C)P(C)}{P(x)}$$

But $P(x)$ is always positive and doesn't depend on C. If we are only looking at what maximizes the posterior, we can safely discard it. Thus, we can make a prediction for the class using only

$$C_{MAP} = \arg\max_{c \in C} P(x|C)P(C)$$

More input features

Why is Naïve Bayes "naïve"

$$C_{MAP} = \arg\max_{c \in C} P(x|C)P(C)$$

What if $\boldsymbol{x}$ is a vector of features with dimensionality D? We'll then have to compute $P(\boldsymbol{x}|c_j)P(c_j))$ for $c_j \in C$ as

$$P(x_1, x_2, \ldots, x_D, c_j) = P(x_1|x_2, \ldots, x_D, c_j)P(x_2| \ldots, x_D, c_j)$$
$$\ldots P(x_{D-1}|x_D, c_j)P(c_j)P(x_D|c_j)P(c_j)$$

If we naïvely assume that the features are conditionally independent given the class, we can simplify this computation to

$$P(x_1, x_2, \ldots, x_D|c_j) = P(x_1|c_j)P(x_2|c_j) \ldots P(x_D|c_j)$$

Not just Gaussian

Gaussian Naïve Bayes

Used with continuous values. We assume that they are normally distributed.

Binomial Naïve Bayes

Features are independent binary variables.

$$P(\boldsymbol{x}|c_j) = \prod_i p_{ji}^{x_i}(1 - p_{ji})^{(1-x_i)}$$

Multinomial Naïve Bayes

Feature vectors represent frequencies of events generated by a multinomial distribution.

References I

Christopher M. Bishop, *Pattern recognition and machine learning (information science and statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

"United States Department of Labor", *Data from national longitudinal youth survey, bureau of labor statistics*, `http://www.bls.gov/nls/nlsy97.htm`, December 1996, Accessed: 21-05-2017.

Q&A

