UNIVERSITY OF BERGEN

INFO 284 - Machine Learning

On probability and Naïve Bayes

Bjørnar Tessem

UNIVERSITY OF BERGEN

Some probability theory

- A random variable is a variable that can take on different values randomly from a set of possible values.
- A random vector is a random variable whose values are vectors.
- Example: if X is a random variable for a day of the week, then can take values from the set

{Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday}

Some probability theory

A discrete random variable is one that has a finite or countably infinite number of possible outcomes

A continuous random variable is associated with a real value (takes values from the set of real numbers)

Probability distribution

- Probability distribution is a function that takes random variable on input and outputs a continuous (real number) value between 0 and 1
- Example: throwing a dice (Outcome = X)
  - $p(X=3) = 1/6$
- Given a random variable x that takes the values from the set $\{x_1, x_2, \ldots x_n\}$, it holds that

$$p(x_1)+p(x_2)+\ldots+p(x_n)=1$$

Probability distribution

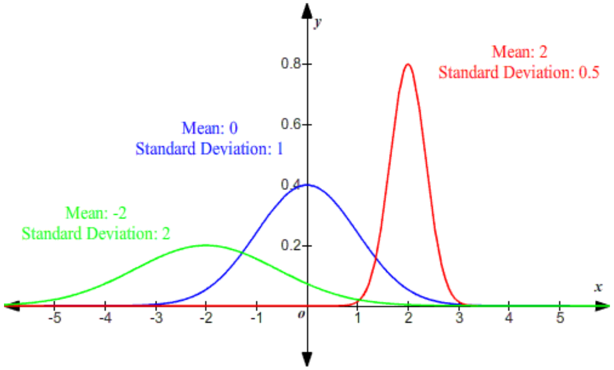
- Probability distribution is a description of how likely a random variable is to take on each of its possible states. Eg.P(x="Monday")
- A probability distribution over many variables is known as a **joint probability distribution**. Eg. P(x="m",y="n")
- Uniform distribution - all values are equally likely Eg. Probability of rolling a 2 on a dice is 1/6.

Probability density

- Continuous random variables represent their probabilities with probability density functions
  - The area under the function is equal to 1.
  - Most famous: Gaussian distribution (Normal distribution)

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

$$\begin{aligned} \mu &= \text{Mean} \\ \sigma &= \text{Standard Deviation} \\ \pi &\approx 3.14159\cdots \\ e &\approx 2.71828\cdots \end{aligned}$$



Conditional probability

- Conditional probability is the probability of some event y, given that some other event x has happened
- Written p(y | x).

$$\boxed{p(y \mid x) = \frac{p(y,x)}{p(x)}}$$

Bayes' rule

- We know $P(a,b) = P(a|b) \cdot P(b) = P(b|a) \cdot P(a)$
- This implies that

$$P(a|b) = \frac{P(b|a) \cdot P(a)}{P(b)}$$

- Assume a is class and b is data
  - Since we use same $P(b)$ for all a, $P(b)$ just helps to normalize
  - Implies: We need to compute $P(b|a) \cdot P(a)$

Independence

- When to variables are independent we have the following relation
  - $P(a \mid b) = P(a)$
  - Example: a dice throw is not dependent on who throws the dice
- Conditional independence:
  - $P(a \mid b, c) = P(a \mid b)$

How well is a machine learning algorithm doing?

- High **accuracy (low generalisation error).**
  - True positives, True negatives
  - False positives, false negatives

- Overfitting - choosing a hypothesis that is overly complex and fits the training data set

- Underfitting - choosing a hypothesis that is simple and allows for too many false positives.

- The complexity of the hypothesis (model) is related to the variability of the input

  - Simple models generalise better to new data.
  - more variation allows for higher complexity

$$P(a|b) = \frac{\boxed{P(b|a)} \cdot \boxed{P(a)}}{P(b)}$$

Naïve Bayes Classifiers

- Apply Bayes rule and independence assumption:
- It is a family of classifiers, all working under the assumption that the value of a particular feature is independent of the value of any other feature
- $\boxed{P(\mathbf{x} \mid y)} = P(x_1, x_2, \ldots, x_n \mid y) = \boxed{P(x_1|y) \cdot P(x_2|y) \ldots P(x_n|y)}$
- The naive Bayes classifier combines a probability model with a decision rule = the function that assigns a class label $\hat{y}$ as follows

$$\hat{y} = \arg\max_y \boxed{P(y)} \boxed{\prod_{i=1}^{n} P(x_i \mid y)},$$

Types of Bayes Classifiers

- Categorical naïve Bayes, Gaussian naïve Bayes, Bernoulli naïve Bayes & Multinomial naïve Bayes
- Categorical naïve Bayes:
  - Use data to create a priori probability distributions and conditional probabilities by counting

$$\hat{y} = \arg\max_y \boxed{P(y)} \boxed{\prod_{i=1}^{n} P(x_i \mid y)},$$

Gaussian Naïve Bayes

- GaussianNB: features have continuous values and we assume that the continuous values associated with each class are distributed according to a Gaussian distribution

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma^2_y}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma^2_y}\right)$$

Multinomial NB

Often applied in document classification

- Use: assume that the features is an integer count of something, i.e. the feature vector is a histogram.

$$p_{yi} = p(x_i|y) = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

- The likelihood of observing a histogram **x** is given by:

$$p(\mathbf{x}|y) = \boxed{\frac{(\sum_i X_i)!}{\prod_i (x_i!)} \prod_i (p_{yi})^{x_i}}$$

$$\hat{y} = \arg\max_y P(y) \boxed{\prod_{i=1}^{n} P(x_i \mid y)},$$

Bernoulli NB

- Bernoulli NB: features are independent booleans describing inputs.
    - Does input have feature or not
- Document classification; If x_i is a boolean expressing the occurrence (x_i=1) or absence (x_i=0) of the i'th term from the vocabulary, then the likelihood of a document given a class C_k is given by

$$p(\mathbf{x}|y) = \prod_i \boxed{p_{yi}^{x_i}}\boxed{(1 - p_{yi})^{(1-x_i)}}$$

Summary of Naive Bayes Classifiers

- Making the statistics:
    - BernoulliNB counts whether a property is present or not. Each feature has value 1 or 0.
    - MultinomialNB takes into account the count of each property for each class. Each feature is a count of property occurences..
    - GaussianNB stores the average value as well as the standard deviation of each feature for each class Each feature is a real number.
- To make a prediction, a data point is compared to the statistics for each of the classes, and the best matching class is predicted.

The alpha parameter

- MultinomialNB and BernoulliNB have a single parameter, alpha, which controls hypothesis complexity
  - alpha “adds” to the data many virtual data points that have positive values for all features
  - This “smoothes” the statistic
  - Large alpha => more smoothing => low hypothesis complexity
  - Setting alpha is not critical for good performance

Use of Naive Bayes Classifiers

- GaussianNB is mainly used for high dimensional data
- MultinomialNB and BernoulliNB are often used for sparse count data, f.ex. Text.
  - Multinomial often best
- Naive Bayes models scale well and are easy to understand
- They may be used with missing data.

UNIVERSITAS BERGENSIS

uib.no