Bayes Classifiers

Lecture # 10

Introduction

The KNN and decision tree classifiers provide definite answers as to whether the data belongs to certain class or not. Their classification could be right or wrong. There are some classifiers that provide a best guess or assign a probability to a dataset to be in a class. Indeed the probability theory forms the basis of many machine learning algorithms. Here we look at the ways the probability theory could be used to classify things. Naïve Bayes classifier is such a technique. It is called “naïve” because its formulation makes some naïve assumptions.

The Baysian learning method calculates explicit probabilities for hypothesis and selects the hypothesis with higher probability. Figure 1 shows datasets with two classes of data. We have a measure of the probability of a new data point (x,y) belonging to class 1, which we call it $p_1(x,y)$, and a probability for the data point belonging to class 2, which we call it $p_2(x,y)$. To classify the data point we use the following rules:

If $p_1(x,y) > p_2(x,y)$ the class is 1

If $p_1(x'y) < p_2(x'y)$ the class is 2.

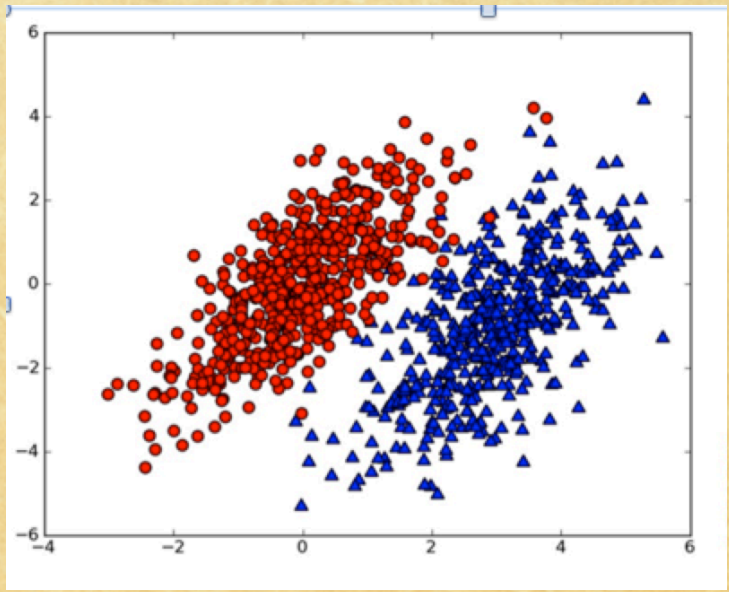
Simply, we choose the class with higher probability to be the class of the data point. This is Bayesian decision theory- choosing the decision with highest probability.

Classification

Probability Distributions

Figure 1 shows two probability distributions with known parameters describing the distributions

Figure 1



The Bayesian Method

The premise of the Bayesian method is that probability statements are not limited to data but can be made for models themselves. Inferences are made by producing Probability Density Functions (PDFs). Model parameters are treated as random variables. Bayesian methods give optimal results given all the available information.

Difference between Classical and Bayesian Approaches

The classical and Bayesian techniques are both concerned with the data likelihood function. In classical statistics the data likelihood function is used to find model parameters that yield the highest data likelihood. The data likelihood cannot be interpreted as a probability density function for model parameters. However, the Bayesian method extends the concept of data likelihood function by adding extra prior information to the analysis and assigning PDFs to all model parameters and models themselves.

The Bayesian method is able to provide a full probabilistic framework for data analysis.

Bayes' Rule

When two continuous random variables are not independent, one could write

$$p(x,y) = p(x|y)\, p(y) = p(y|x)\, p(x)$$

Where p(x|y) and p(y|x) are conditional probabilities and p(x) and p(y) are marginal probabilities. The marginal probabilities are defined as

$$p(x) = \int p(x,y)dy$$

$$p(x) = \int p(x|y)p(y)dy$$

Complete knowledge of marginal probability p(y) and conditional probability p(y|x) are needed to reconstruct p(x,y). The continuous version of the law of probability then becomes

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

The Bayes' rule is then derived from the above equations

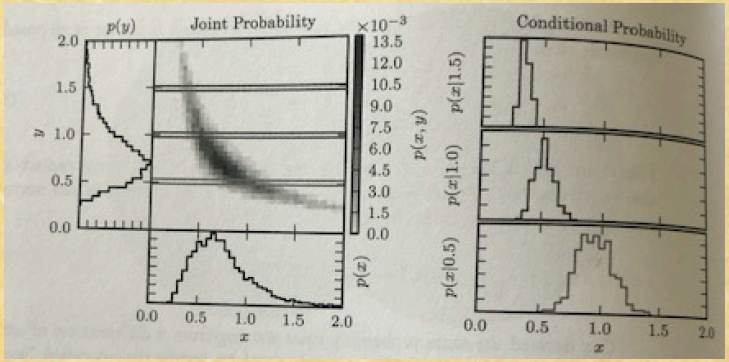
Bayes rule connects conditional and marginal probabilities to each other. In the case of a discrete random variable, y_i , with M possible values, the integral becomes

$$p(y_i|x) = \frac{p(x|y_i)p(y_i)}{p(x)} = \frac{p(x|y_i)p(y_i)}{\sum_{j=1}^{M} p(x|y_i)p(y_i)}$$

Joint Conditional Probability

Figure 2: Joint and Conditional probability distributions

Two dimensional probability distribution in Figure 2 showing $p(x,y)$. The two panels on the left and bottom show marginal distributions in x and y. The three panels on the right show conditional probability distributions $p(x|y)$ for three different values of y, as marked on the left panel.



Bayes' Theorem

By applying Bayes' rule to the likelihood function p (D|M), one obtains the Bayes' theorem

$$p\,(M|D) = \frac{p(D|M)\,p(M)}{P(D)}$$

Where D stands for data and M stands for model. The Bayes' theorem combines an "initial belief" with new data and arrives at an "improved belief". The "improved belief" is proportional to the product of the "initial belief" and the probability that initial belief generated the observed data.

General Expression of The Bayes' Theorem

Here we define the presence of a prior information, I and that models are mostly described by parameters whose values we need to estimate from data:

$$p(M,\theta|D,I) = \frac{p(D|M,\theta,I)p(M,\theta|I)}{p(D|I)}$$

Model M includes k model parameters θ_p, p=1,…k shown as vector θ with components θ_p. The result p(M, q |D, I) is called the posterior PDF for model M and parameters θ, given data D and other prior information I. The term p(D|, M, θ, I) is the likelihood of data given some model M and some fixed values of parameter θ describing it and the prior information I.

The term p(M, θ| I) is the joint probability for model M and its parameters θ in the absence of any of the data used to compute likelihood. This is simply called the prior. The prior can be expanded as:

P(M, q| I) = p (θ |M, I) p(M |I)

We only need to specify p (θ | M, I). The term p (D|I) is the probability of data or the prior predictive probability for D

Explaining the Bayesian Approach

The Bayesian approach can be considered as formalizing the process of continually refining our state of knowledge about the world, beginning with no data (as depicted by the prior), then updating that by multiplying in the likelihood once the data D are observed to obtain the posterior. When more data are taken , then the posterior based on the first data set can be used as the prior for the second analysis.

Prior: A prior incorporates all other knowledge that might exist, but is not used when computing the likelihood $p(D \mid M, \theta, I)$.

(from the book on Statistics, Data Mining and Machine Learning in Astronomy by Z. Ivezic, A. Connolly..)

Bayesian Model Selection

Bayes's theorem calculates the posterior PDF of parameters describing a single model, with the model assumed to be true. In model selection and hypothesis testing we often come up with the question as which model is best supported by the available data? For example, we may ask as set of data $[x_i]$ is better described by a Gaussian or a Poisson distribution? Or if a set of points is better described by a straight line or a parabola?

To find out which of the two models M_1 or M_2 is best supported by the data, we compute the odds ratio of model M_2 over model M_1 as

$$Q_{21} = \frac{P(M_2|D,I)}{P(M_1|D,I)}$$

The posterior probability for model M (M_1 or M_2) given data D, $p(M|D, I)$ in this expression can be obtained from the posterior PDF $p(M, \theta|D, I)$ using marginalization integration over the model parameter space spanned by θ. The posterior probability that model M is correct given data D (a number between 0 and 1) is

$$p(M|D,I) = \frac{p(D|M,I)p(M|I)}{p(D|I)}$$

Where

$$E(M) \equiv p\,(D|M,I) = \int p(D|M,\theta,I)\,p(\theta\,|M,I)\,d\theta$$

Which is called the marginal likelihood for model M and it quantifies the probability that the data D would be observed if model M were the correct model. Therefore

$$Q_{21} = \frac{E(M_2)p(M_2,I)}{E(M_1)p(M_1,I)} = B_{21}\frac{P(M_2|I)}{P(M_1|I)}$$

Where $B_{21} = E(M_2)/E(M_1)$ is called the Bayes factor.

To interpret the Q_{21} values, one may assume the odds ratio of $Q_{21} > 10$ to imply strong evidence in favor of M_2 (M_2 is 10 times more probable than M_1) and $Q_{21} > 100$ is decisive evidence that M2 is the right model for the data.

Summary of Bayes Probability

Bayes Theory:

In ML we are often interested in determining the best hypothesis from some space H, given the observed training data D. By the "best hypothesis" here we mean the most probable hypothesis given the data D and any prior knowledge about prior probabilities of various hypothesis in H. Bayes theorem provides a direct method for calculating the probability of a hypothesis based on its prior probability.

We use the following notations. We shall write p(h) as the initial probability that hypothesis h holds before we have observed training data. p(h) is called the prior probability of h and contains any background knowledge we may have about the chance that h is a correct hypothesis. Similarly, p(D) denotes the prior probability that training data D will be observed (the probability of D given no knowledge about which hypothesis holds). Next we define conditional probability p(D|h) which is the probability of observing data D given some world in which hypothesis h holds. In other words, we write p(x|y) to denote the probability of x given y. In ML we are mainly interested in the probability p(h | D) that h holds given the observed training data D. p(h|D) is called the posterior probability of h.

Bayes theorem is used in Bayesian learning methods because it provides a way to calculate posterior probability p(h|D), from the prior probability p(h), together with p(D) and p(D|h). It states that

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

Here p(h | D) increases with p(h) and with p(D|h) according to Bayes theorem. It is also reasonable to see that p(h|D) decreases as p(D) increases because the more probable it is that D will be observed independent of h, the less evidence D provides in support of h.

Maximum Likelihood

In many learning scenarios one considers some set of hypothesis H and needs to find the most probable hypothesis h $\in$ H given the observed data D (or one of the maximally probable if there are many). We can determine this using Bayes theorem to calculate posterior probability of each candidate hypothesis. Any such maximally probable hypothesis is called maximum a posteriori (MAP)

$$h_{MAP} = argmax\, p(h|D) = argmax\, \frac{p(D|h)p(h)}{p(D)} = argmax\, p(D|h)p(h)$$

Here p(D) is dropped from the last step because it is a constant independent of h.

In some cases, we will assume that every hypothesis in H is equally probable a priori ($p(h_i) = p(h_j)$ for all h_i and h_j in H). In this case we could further simplify the above equation and need only use p(D|h). p(D|h) is often called the likelihood of the data D given h and any hypothesis that maximizes p(D|h) is called maximum likelihood

$H_{MaxL} = argmax\, p\,(D \mid h)$

Example # 1: Conditional Probability

Let's assume we have a jar containing seven stones. Three of these stones are grey and four are black. The chance of randomly selecting a grey stone is *p(grey)=3/7* while selecting a black stone is *p(black)= 4/7*. Now, we divide these into two buckets. What is the probability of drawing a grey stone from bucket B? This is known as conditional probability. We are calculating the probability of a grey stone, given that the unknown stone is coming from bucket B. We can write this as *P(grey |bucket B)*, which is read as "the probability of grey given bucket B". It is easy to find *p(grey | bucket A)= 2/4* and *p(grey | bucket B)= 1/3*. In other words, we can say

$$p(grey|bucket\ B) = \frac{p(grey\ and\ bucket\ B)}{p(bucket\ B)}$$

Expression of Bayes Probability

Which is calculated as

$$p(grey\ and\ bucket\ B) = \frac{1}{7}\ ;\ p(bucket\ B) = \frac{3}{7}$$

Therefore

$$p(grey|bucket\ B) = \frac{1/7}{3/7} = \frac{1}{3}$$

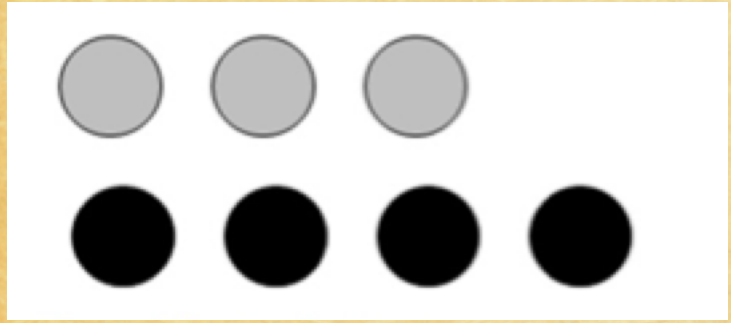
Bayes rule tells us how to swap the symbols in a conditional probability statement. If we have $p(x|c)$ but want to have $p(c|x)$, we use the following formula

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)}$$

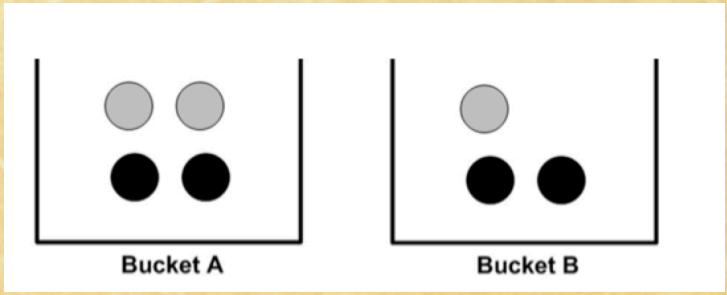
which is just another way of interpreting the relation we already came to.

Example

Collection of seven stones that are grey and black. If we randomly select a stone from this set, the probability that it will be grey or black stone is 3/7 and 4/7 respectively



Seven stones divided in two buckets



Example # 2: Bayes Classifiers

Consider a medical diagnosis problem in which there are two alternative hypothesis: (1). The patient has a particular form of cancer (2). The patient does not. There are two possible outcomes from the test:

$\oplus$ *positive*) and $\ominus$ (*negative*).

We have prior knowledge that over the entire population of people, only 0.008 have this disease. There are some uncertainties in the lab results.

$$p(cancer) = 0.008 \quad p(not\ cancer) = 0.992$$
$$p(\oplus \mid cancer) = 0.98 \quad p(\ominus \mid cancer) = 0.02$$
$$p(\oplus \mid not\ cancer) = 0.03 \quad p(\ominus \mid not\ cancer) = 0.97$$

The lab returns a correct positive result in only 98% of cases in which the disease is actually present. In other cases the test returns the opposite result. The results can be summarized as follows:

This is our training set. Suppose now we observe a new patient for whom the lab has returned a positive result. Should we diagnose the patient as having cancer or not? From the above relation

$$h_{MaxL} = argmax\ p(D|h)$$

We can calculate

$$p\ (\oplus\ |cancer)p(cancer) = (0.98)(0.008) = 0.0078$$
$$p((\oplus\ |no\ cancer)p(no\ cancer) = (0.03)(0.992) = 0.0298$$

Therefore $h_{MAP} = no\ cancer$. Here we use the property that

$$p(cancer|\ \oplus) + p(no\ cancer|\ \oplus) = 1$$

Either the patient has cancer or does not.

Note that here the hypothesis is not completely accepted or rejected but a probability is assigned to it

Naïve Bayes Classifiers

The naïve Bayes classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function f(x) can take on any value from some finite set V. The Bayesian approach to classifying the new instance is to assign the most probable target value, v_{MAP}, given the attribute values (a_1, a_2,…a_n) that describe the instance

$$v_{MAP} = argmax\ P(v_j|a_1, a_2, \dots a_n)$$

This can be written using Bayes theory

$$v_{MPA} = argmax\ \frac{p(a_1, a_2 \dots a_n|v_j)p(v_j)}{p(a_1, a_2 \dots a_n)}$$
$$= argmax\ p(a_1, a_2 \dots a_n|v_j)p(v_j)$$

We could estimate the two terms in the above equation based on the training data. Each of the $p(v_j)$ values is estimated by counting the frequency with which each target value vj occurs in the training data. However, estimating different $p(a_1, a_2 \ldots, a_n)$ terms in this fashion is not feasible unless we have a very large training dataset. The problem is that the number of these terms equals the number of possible instances times the number of possible target values. Therefore, we need to see every instance in the instance space many times in order to obtain reliable estimates (this becomes clear shortly after we go through the example).

The naïve Bayes classifier is based on the simplifying assumption that the attribute values are conditionally independent. In other words, given the target value of the instance, the probability of observing the conjunction $a_1, a_2 \ldots, a_n$ is just the product of the probabilities of individual attributes:

$$p(a_1, a_2, \ldots a_n | v_j) = p(a_1|v_j)\, p(a_2|v_j) \ldots p(a_n|v_j) = \prod_i p(a_i|v_j)$$

The Naïve Bayes classifier then becomes

$$v_{NB} = argmax\, p(v_j) \prod_i p(a_i|v_j)$$

Where v_{NB} is the target value output by naïve Bayes classifier. In naïve Bayes classifier the number of distinct $p(a_i \mid v_j)$ terms that must be estimated from the training data is just the number of distinct attribute values times the number of distinct target values.

In summary the naïve Bayes learning method involves a learning step in which the various $p(v_j)$ and $p(a_i \mid v_j)$ terms are estimated based on their frequency over the training data. The set of these estimates corresponds to the learned hypothesis. This hypothesis is then used to classify each new instance by applying the above relation.

Example # 3: Naïve Bayes

Consider data in the following Table (taken from Machine Learning book by T. Mitchel). Here the target attribute is PlayTennis which can have values yes or no for different Saturday mornings. This is to be predicted based on other attributes of the morning in question. We now apply the naïve Bayes classifier to concept learning. The table provides a sample of 14 training examples of the target concept "PlayTennis" when each day is described by attributes outlook, Temperature, Humidity and wind. Here we use the naïve Bayes classifier and the training data from this table to classify the following instance:

(Outlook=sunny, Temperature=cool, Humidity=high, wind= strong)

Our task is to predict the target value (yes or no) of the target concept PlayTennis for the new instance. Using the above relation

$$v_{NB} = argmax\, p(v_j) \prod_i p(a_i|v_j)$$
$$= argmax\, p(v_j) p(Outlook = sunny|v_j) p(Temperature = cool|v_j) p(Humidity = high|v_j)$$
$$p(Wind = strong|v_j)$$

Here a_i is instantiated to have the attributes of the day in question (the new instance). To calculate v_{NB} we now need 10 probabilities that can be estimated from the training dataset. First the probability of different target values can easily be estimated based on their frequencies over the 14 training examples

$$p(PlayTennis = yes) = \frac{9}{14} = 0.64$$
$$p(PlayTennis = no) = \frac{5}{14} = 0.36$$

Similarly we can estimate the conditional probabilities. For example, those for “wind=strong” are

$$p(Wind = strong \mid PlayTennis = yes) = \frac{3}{9} = 0.33$$
$$p(Wind = strog \mid PlayTennis = no) = \frac{3}{5} = 0.60$$

We use these probability strengths and similar probabilities for other attributes. We then calculate v_{NB} as follows:

$$p(yes)p(sunny|yes)p(cool|yes)p(high|yes)P(strong|yes) = 0.0053$$
$$p(no)p(sunny|no)p(cool|no)p(high|no)p(strong|no) = 0.0206$$

Comparing the two probabilities the naïve Bayes classifier assigns PlayTennis=no to this new instance based on the training data. The conditional probability for the current example (considering that the sum of two probabilities should be one) is

$$\frac{0.206}{0.206 + 0.053} = 0.795$$

Training examples for the target concept PlayTennis

Day					
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					No

Sources used for this lecture

Machine Learning In Action

By Peter Harrington

Statistics, Data Mining, and Machine Learining in Astronomy

Z. Ivezic, A. Connolly, J. VanderPlas & A. Gray