Maximum likelihood estimation

Jyotirmoy Banerjee

May 26, 2020

1 Introduction

Probability mass function: A discrete random variable is a random variable whose range is finite or countably infinite. The probability mass function of a discrete random variable X is

$$f_X(x) = P\{X = x\}$$

The mass function has two basic properties:

- $f_X(x) \geq 0$ for all x in the state space
- $\sum_x f_X(x) = 1$

Probability density function: Let X be a random variable whose cumulative distribution function F_X has a derivative. The function f_X satisfying

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

is called the probability density function and X is called a continuous random variable. By the fundamental theorem of calculus, $F_X'(x) = f_X(x)$. We can compute compute probabilities using

$$P\{a < X \leq b\} = F_X(b) - F_X(a) = \int_a^b f_X(t)\, dt$$

Independence: X and Y are independent if:

$$P(X, Y) = P(X)P(Y)$$
$$\text{Joint} = \text{Marginal} \times \text{Marginal}$$

Conditional probability: Probability of X given that Y happened:

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$
$$\text{Conditional} = \frac{\text{Joint}}{\text{Marginal}}$$

Chain rule:

$$\begin{aligned} P(X,Y) &= P(X|Y)P(Y) \\ P(X,Y,Z) &= P(X|Y,Z)P(Y|Z)P(Z) \\ P(X_1,X_2,\ldots,X_N) &= \prod_{i=1}^{N} P(X_i|X_1,X_2,\ldots,X_{i-1}) \end{aligned}$$

Sum rule: Marginalization

$$p(X) = \int_{-\infty}^{\infty} p(X,Y)\, dY$$

Bayes theorem

- θ - Parameters
- X - Observations

$$\begin{aligned} P(\theta|X) &= \frac{P(\theta,X)}{P(X)} \\ &= \frac{P(X|\theta)P(\theta)}{P(X)} \\ \text{Posterior} &= \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} \\ &\propto \text{Likelihood} \times \text{Prior} \end{aligned}$$

2 Likelihood

Given a statistical probability mass function or density, say $f(x,\theta)$, where θ is an unknown parameter, the *likelihood* is f viewed as a function of θ for a fixed, observed value of x of the random variable X. The probability density function is the continuous analogue of probability mass function.

Given the outcome x of the random variable X, the likelihood function, which is a function of θ, is given as:

$$\mathcal{L}(\theta|x) = f_\theta(x) = P_\theta(X = x) = P(X = x \mid \theta)$$

For example -

1. Suppose we flip a coin with success probability of θ
2. Recall that the mass function for x

$$f(x,\theta) = \theta^x(1-\theta)^{1-x} \quad \text{for} \quad \theta \in [0,1]$$

where x is either 0 (tails) or 1 (heads)

3. Suppose that the result is a head. The likelihood is

$$\mathcal{L}(\theta|1) = \theta^1(1-\theta)^{1-1} = \theta \quad \text{for} \quad \theta \in [0,1]$$

4. Therefore, $\mathcal{L}(0.5|1)/\mathcal{L}(0.25|1) = 2$
5. There is twice as much evidence supporting the hypothesis that $\theta = 0.5$ to the hypothesis that $\theta = 0.25$

3 Maximum likelihood and Maximum a posteriori

Maximum a posteriori (MAP) and maximum likelihood estimation (MLE) are stated as follows

$$\begin{aligned}\theta_{\mathcal{MAP}} &= \underset{\theta}{\operatorname{argmax}}\, P(\theta|x) \\ &= \underset{\theta}{\operatorname{argmax}}\, P(x|\theta)P(\theta) \\ \theta_{\mathcal{MLE}} &= \underset{\theta}{\operatorname{argmax}}\, P(x|\theta)\end{aligned}$$

where x is the input data and θ is the output parameter. $\mathcal{L}_\theta = P(x|\theta)$.

4 Maximum likelihood

- Let's suppose we have observed 10 data points from some process. For example, each data point could represent the length of time in seconds that it takes a student to answer a specific exam question.
- We assume that they have been generated from a process that is adequately described by a Gaussian distribution. How do we calculate the maximum likelihood estimates of the parameter values of the Gaussian distribution μ and σ?
- Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed.
- What we want to calculate is the total probability of observing all of the data, i.e. the joint probability distribution of all observed data points. To do this we would need to calculate some conditional probabilities, which can get very difficult. So it is here that we will make our first assumption. The assumption is that each data point is generated independently of the others. This assumption makes the maths much easier.
- If the events (i.e. the process that generates the data) are independent, then the total probability of observing all of data is the product of observing each data point individually (i.e. the product of the marginal probabilities).

The probability density of observing a single data point x, that is generated from a Gaussian distribution is given by:

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The semi colon used in the notation $P(x;\mu,\sigma)$ is there to emphasise that the symbols that appear after it are parameters of the probability distribution. So it should not be confused with a conditional probability (which is typically represented with a vertical line e.g. $P(A|B)$).

The total (joint) probability density of observing n data points is given by:

$$P(x_1, x_2, \cdots, x_n;\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

We just have to figure out the values of μ and σ that results in giving the maximum value of the above expression. This expression can be differentiated to find the maximum.

The above expression for the total probability is difficult to differentiate, so it is almost always simplified by taking the natural logarithm of the expression. This is absolutely fine because the natural logarithm is a monotonically increasing function. This means that if the value on the x-axis increases, the value on the y-axis also increases. This is important because it ensures that the maximum value of the log of the probability occurs at the same point as the original probability function. Therefore we can work with the simpler log-likelihood instead of the original likelihood.

5 Bayes or Maximum likelihood classifier

The maximum likelihood classifier is one of the most popular methods of classification, in which x with the maximum likelihood is classified into the corresponding class. The likelihood L_k is defined as the posterior probability of a pixel belonging to class k.

$$L_k = P(k|x) = \frac{P(x|k)P(k)}{P(x)} = \frac{P(x|k)P(k)}{\sum P(x|i)P(i)}$$

where $P(k)$ is the probability of class k.

Usually $P(k)$ are assumed to be equal to each other and $P(i) * P(x|i)$ is also common to all classes. Therefore L_k depends on $P(x|k)$ or the probability density function.

For mathematical reasons, a multivariate normal distribution is applied as the probability density function. In the case of normal distributions, the likelihood can be expressed as follows.

$$L_k = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

6 Naïve Bayes

The Naïve Bayes classifier is an approximation to the Bayes classifier, in which we assume that the features are conditionally independent given the class instead of modelling their full conditional distribution given the class.

$$\begin{aligned}
P(k|d) &= P(k|x_1, x_2, \cdots, x_n) \\
P(k|x_1, x_2, \cdots, x_n) &\propto P(x_1, x_2, x_3, \cdots, x_n, k) \\
&= P(x_1|x_2, x_3, \cdots, x_n, k)P(x_2, x_3, \cdots, x_n, k) \\
&= P(x_1|x_2, x_3, \cdots, x_n, k)P(x_2|x_3, \cdots, x_n, k)P(x_3, \cdots, x_n, k) \\
&= P(x_1|x_2, x_3, \cdots, x_n, k)P(x_2|x_3, \cdots, x_n, k) \cdots P(x_n|k)P(k) \\
&= P(k)P(x_1|k)P(x_2|k) \cdots P(x_n|k) \\
&= P(k)\prod_{i=1}^{n} P(x_i|k)
\end{aligned}$$

where d is the document and x_i is the term occurring in the document.

$$\begin{aligned}
P(k|d) &= P(k|x_1, x_2, \cdots, x_n) \\
&= P(k)\prod_{i=1}^{n} P(x_i|k)
\end{aligned}$$

For the prior the estimate is $P(k) = \frac{N_k}{N}$, where N_k is the number of documents in class k and N is the total number of documents.

The conditional probability $P(x_i|k)$ is estimated as the relative frequency of term x_i in documents belonging to class k:

$$P(x_i|k) = \frac{T_{kx_i}}{\sum_{x_{i'} \in V} T_{kx_{i'}}}$$

where T_{ki} is the number of occurrence of x_i in training documents from class k, including multiple occurrences of a term in a document. V is the vocabulary.

6.1 Laplace smoothing

If a given class and feature value never occur together in the training data, then the frequency-based probability estimate will be zero, because the probability estimate is directly proportional to the number of occurrences of a feature's value. This is problematic because it will wipe out all information in the other probabilities when they are multiplied. Therefore, it is often desirable to incorporate a small-sample correction, called pseudocount, in all probability estimates such that no probability is ever set to be exactly zero. This way of regularizing naïve Bayes is called Laplace smoothing when the pseudocount is one, and Lidstone smoothing in the general case. Laplace smoothing is given as:

$$P(x_i|k) = \frac{T_{ki} + 1}{\sum_{x_{i'} \in V}(T_{kx_{i'}} + 1)}$$

6.2 Variants of Multinomial Naïve Bayes

Let $\mathrm{tf}_{(x_i,d)}$ be the term frequency of x_i in document d. Then $P(k|d)$ is computed as follows:

$$\begin{aligned} P(k|d) &= P(k|x_1, x_2, \cdots, x_n) \\ &= P(k) \prod_{i=1}^{n} P(x_i|k)^{\mathrm{tf}_{(x_i,d)}} \end{aligned}$$

In the context of document classification and possible ways to alleviate those problems is by including the use of tf-idf weights instead of raw term frequencies and document length normalization, to produce a naïve Bayes classifier that is competitive with support vector machines.

7 Maximum likelihood regression

The distribution of X is arbitrary. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some parameters β_0 and β_1, and some random noise variable ϵ. The noise is independent of X.

Because of these stronger assumptions, the model tells us the conditional pdf of Y for each x, $P(y|X = x; \beta_0, \beta_1, \sigma)$. Given any data set $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$, we can now write down the probability density, under the model, of seeing that data:

$$\prod_{i=1}^{n} P(y_i|x_i; \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i + \epsilon))^2}{2\sigma^2}\right)$$

In multiplying together the probabilities like this, we are using the independence of the Y_i. This is the likelihood function. In the method of maximum likelihood, we pick the parameter values which maximize the likelihood, or, equivalently, maximize the log-likelihood.

8 Logistic regression using maximum likelihood formulation

Assume that $P(Y = 1|X = x) = p(x; \theta)$, for some function parametrized by θ. Further assume that observations are independent of each other. Then the (conditional) likelihood function is:

$$\prod_{i=1}^{n} P(Y = y_i|X = x_i) = \prod_{i=1}^{n} p(x_i; \theta)^{y_i}(1 - p(x_i; \theta)^{1-y_i})$$

8.1 logistic function

In logistic regression, we use the logistic function:

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

8.2 odds

After a bit of manipulation of, we find that:

$$\frac{p(X)}{1 - p(X)} = \exp(\beta_0 + \beta_1 X)$$

The quantity $p(X)/[1p(X)]$ is called the odds, and can take on any value between 0 and inf. Values of the odds close to 0 and inf indicate very low and very high probabilities of default, respectively.

8.3 log-odds or logit

By taking the logarithm of both sides of, we arrive at:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

The left-hand side is called the log-odds or logit. We see that the logistic regression model has a logit that is linear in X.

8.4 likelihood

The likelihood function is then:

$$L(\hat{\beta}_0, \hat{\beta}_1) = \prod_{i=1}^{n} p(x_i; \theta)^{y_i} (1 - p(x_i; \theta)^{1-y_i})$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize this likelihood function.

8.5 Newton's Method for Numerical Optimization

There are a number of methods for numerical optimization; one of the most ancient yet important of them is Newton's method (alias "Newton-Raphson").

Consider the simplest case of minimizing a function of one scalar variable, say $f(\beta)$. We want to find the location of the global minimum, β^*. We suppose that f is smooth, and that β^* is a regular interior minimum, meaning that the derivative at β^* is zero and the second derivative is positive. Near the minimum we could make a Taylor expansion:

$$f(\beta) \approx f(\beta^*) + \frac{1}{2}(\beta - \beta^*)^2 \frac{d^2 f}{d\beta^2}\Big|_{\beta=\beta^*}$$

(We can see here that the second derivative has to be positive to ensure that $f(\beta) > f(\beta^*)$.) In words, $f(\beta)$ is close to quadratic near the minimum.

Newton's method uses this fact, and minimizes a quadratic approximation to the function we are really interested in. (In other words, Newton's method is to replace the problem we want to solve, with a problem which we can solve.) Guess an initial point β^0. If this is close to the minimum, we can take a second order Taylor expansion around β^0 and it will still be accurate:

$$f(\beta) \approx f(\beta^0) + (\beta - \beta^0)\frac{df}{d\beta}\Big|_{\beta=\beta^0} + \frac{1}{2}(\beta - \beta^0)^2\frac{d^2f}{d\beta^2}\Big|_{\beta=\beta^0}$$

Now it is easy to minimize the right-hand side of equation. We just take the derivative with respect to β, and set it equal to zero at a point we will call β^1:

$$\begin{aligned} 0 &= f'(\beta^0) + (\beta^1 - \beta^0)f''(\beta^0) \\ \beta^1 &= \beta^0 - \frac{f'(\beta^0)}{f''(\beta^0)} \end{aligned}$$

The value β^1 should be a better guess at the minimum β than the initial one β^0 was. So if we use it to make a quadratic approximation to f, we will get a better approximation, and so we can iterate this procedure, minimizing one approximation and then using that to get a new approximation:

$$\beta^{(n+1)} = \beta^{(n)} - \frac{f'(\beta^{(n)})}{f''(\beta^{(n)})}$$