Bayesian Applied Methods

Student

Máster en Ciencia de Datos
Universidad Autónoma de Madrid

tu-web.es

Contents

Part I. Introduction

1. Introduction

Consider a set of variables $X_1, \ldots, X_n$, we will define a model that will consider the joint probability density function of the variables

$$P(X_1, \ldots, X_n)$$

There are a few forms of representing this

- Bayesian Networks, which are represented by directed graphs in which each node represents conditional probabilities, and the edges represent dependencies. (Drawing from the blackboard)
- Markov Networks. In this case, we have undirected graphs, where the edges are *factors* (tables of probability)

Bayesian networks have the *factors* in the nodes, while Markov networks have the factors in the edges.

Our goal will be to make inference about variables using the available information. Sometimes, making inference is understood as **marginalizing the joint distribution**.

$$P(A) = \sum_{B,C} P(A, B, C) = \sum_{B,C} P(A)P(B|A)P(C|B, A)$$

There are different algorithms to compute this probabilities, such as *variable elimination* or *message-passing*.

Also, there are different ways of **reasoning**, such as

- Causal reasoning, which studies causalities
- Evidential reasoning

With all this types of reasoning, we are assuming **two main points**:

- That we have all the information about both the structure of the network and the *factors* or probabilities.

Having both the network structure and the probabilities of each of the factors, we can marginalize to obtain the joint probability distributions. However, we can consider a case where we do **not** know the structure of the network but we know the probabilities of the factors. This is also a *branch* of study, which we will not go deep in. A last case is the one where we **know** the structure of the network, but we are **lacking** parts of the table which we would like to infer. This is the case that we will focus in this course.

Example 1.1. Consider the following random variable, modeling the probability of obtaining heads or tails in a coin flip

$$v^n = \begin{cases} 1 & \text{heads} \\ 0 & \text{tails} \end{cases}$$

However, our coin might have different weights for each of the outcomes (biased coin). For instance, consider that $\theta = P(\text{heads}) = P(v^n = 1|\theta)$. Hence, $P(\text{tails}) = 1 - \theta$. In this case, our goal would be to **determine** θ.

If we tossed the coin n times, and **we knew the probability** θ, the coin tosses would be **independent**. However, if we **do not know** the probability θ, the coin tosses **would be dependent** since the **outcome** of the experiment affects θ. (Diagrams from the blackboard). In this case, the joint probability would be

$$P(\theta, v^1, \dots, v^n) = P(\theta)\prod_{i=1}^{n} P(v^i|\theta)$$

End of example.

Estimations of the parameters are sometimes done using the empirical distribution function. However, there are other methods of estimating the joint pdf, such as *maximum likelihood estimation*.

1.1. Maximum likelihood estimation

We define the **likelihood** of the data as

$$L(\theta; D) = P(D|\theta) = \prod_{j=1}^{N} P(v^j|\theta)$$

Maximum likelihood estimation determines the likelihood function and tries to find (or approximate) a maximum of it.

We can generalize our previous coin toss example. Consider that we obtained M_h heads and M_t tails in our coin toss problem. In this case, our likelihood function would be

$$L(\theta; M_h, M_t) = \theta^{M_h}(1-\theta)^{M_t}$$

The most common approach is to apply the logarithm to the likelihood function, which is a monotonous increasing function, to convert the product into sums, and then maximize the **log-likelihood**. In the previous example, our log-likelihood would be

$$\ell = \log L(\theta; M_h, M_t) = M_h \log\theta + M_t \log(1-\theta)$$

Consider the case of a three node bayesian network, with joint pdf:

$$P(A,B,Y) = P(A)P(B|A)P(Y|A,B).$$

We would like to obtain the probabilities in a table in each case. Extracting data is to observe (N) realizations of an experiment. We would like to use this data to determine $\theta_A, \theta_B, \theta_{Y|A,B}$. Recall that these θs **are not distribution parameters but computed probabilities of observations**. We estimate this parameter set:

$$\Theta = \{\theta_A, \theta_B, \theta_{Y|A,B}\}$$

considering that we have to compute θ_a for all $a \in Values(A)$, $theta_b$ for all $b \in Values(B)$ and $\theta_{y|a,b}$ for all $a \in Values(A), b \in Values(B)$ and $y \in Values(Y)$

Let us compute the likelihood of this Θ:

$$\begin{aligned}
L(\Theta, D) &= P(D|\Theta) \\
&= \prod_{j=1}^{N} P(a[j], b[j], y[j]|\Theta) \\
&= \prod_{j=1}^{N} P(a[j]|\Theta)P(b[j]|\Theta)P(y[j]|a[j], b[j], \Theta) \\
&= \prod_{j=1}^{N} P(a[j]|\theta_A)P(b[j]|\theta_B)P(y[j]|a[j], b[j], \theta_{Y|A,B}) \\
&= \prod_{j=1}^{N} L(\theta_A, D)L(\theta_B, D)L(\theta_{Y|A,B}, D)
\end{aligned}$$

So we have expressed the likelihood of the parameters Θ as the product of the likelihood of the individual parameters θ_i. We can extend this to a more general case.

> **Proposition 1.1.** Let $X_1, \ldots, X_K$ be random variables with a bayesian network dependence. Let D be a sample. Let $\tilde{U} = \mathrm{Par}_g(X_i)$. Then,
>
> $$\begin{aligned}
L(\Theta, D) &= \prod_{j=1}^{N} P(\tilde{x}[j]|\Theta) \\
&= \prod_{j=1}^{N}\prod_{i=1}^{K} P(x[j]|\tilde{u}_i[j], \Theta_i) \\
&= \prod_{i=1}^{K} L(\Theta_i|D)
\end{aligned}$$

Proposition 1.2. The MLE of the general case of a bayesian network is given by:

$$\Theta_{x|\tilde{u}} = \frac{M[X = x, \tilde{U} = \tilde{u}]}{M[\tilde{U} = \tilde{u}]}$$

where M is the counting function.

1.2. Limitations of the frequentist approach

Frequentist maximum likelihood estimation has a few limitations:

1. We may assign zero probability to events just because they do not appear in the considered sample. When the number of observations is low, the probability estimation is very poor and some possible values may even not appear.
2. Fragmentation: The number of values to estimate de CDP table increases exponentially with the number of parents $|\tilde{U}|$. A common solution is to represent the data in a smaller number of states.
3. Overfitting

Example: spam detection using KNN.

INCLUDE IMAGE!!!

Consider a sample of $1M$ examples. If we had $K = 30$ nodes, we would have to estimate 2^{30} parameters. This is clearly not feasible.

We want to solve this using **bayesian estimation**.

2. The bayesian approach

Until now, we have been dealing with the likelihood of our data $L(\Theta; D)$. In the bayesian approach, we try to compute the **posterior probability** using the Bayes' theorem:

$$P(\Theta|D) = \frac{P(D|\Theta)P(\Theta)}{P(D)}$$

where $P(\Theta)$ is the prior probability of our parameter, $P(D)$ is the probability of the data, and $P(D|\Theta)$ is the likelihood.

To achieve this, we **need a good estimation of the prior probability** $P(\Theta)$. We will know study a few distributions that can be used as prior distributions:

2.1. Beta and Dirichlet distributions

This is the univariate version of the Dirichlet distribution. It is given by:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}$$

is essentially a normalizing factor.

We generalize this formula to obtain the Dirichlet distribution. (Complete from jupyter notebook)

Why is this distribution interesting?

> **Definition 2.1.** If the posterior distribution $P(\theta|D)$ is in the same probability distribution family as the prior probability distribution $P(\theta)$, then the prior and the posterior are called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood function $P(D|\Theta)$.

Let us see this in an example using the beta distribution in the biased coin example:

$$P(\Theta|D) = \frac{\theta^{M_h}(1-\theta)^{M_t} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}}{P(D)} \frac{1}{B(\alpha, \beta)} = \frac{1}{P(D)B(\alpha, \beta)} \theta^{M_h+\alpha-1}(1-\theta)^{M_t+\beta-1}$$

Let us compute the probability of the data

$$\begin{aligned}
P(D) &= \int_0^1 P(D|\theta)P(\theta)d\theta \\
&= \frac{1}{B(\alpha,\beta)}\int_0^1 \theta^{M_h+\alpha-1}(1-\theta)^{M_t+\beta-1}d\theta \\
&= \frac{1}{B(\alpha,\beta)}\frac{\Gamma(M_h+\alpha)\Gamma(M_t+\beta)}{\Gamma(M_h+M_t+\alpha+\beta)}
\end{aligned}$$

We use it back in the previous expression

$$\begin{aligned}
\frac{1}{P(D)B(\alpha,\beta)}\theta^{M_h+\alpha-1}(1-\theta)^{M_t+\beta-1} &= \frac{\Gamma(M_h+\alpha)\Gamma(M_t+\beta)}{\Gamma(M_h+M_t+\alpha+\beta)}\theta^{M_h+\alpha-1}(1-\theta)^{M_t+\beta-1} \\
&= B(\theta; M_h+\alpha, M_t+\beta)
\end{aligned}$$

2.2. Prediction using Bayesian Networks

Consider that we have observed n random variables $v[1],\dots,v[n]$, we would like to predict the following value $v[n+1]$. This can be expressed as:

$$P(v[n+1]=1|v[1],\dots,v[n]) = \int_0^1 P(v[n+1]=1|\theta)P(\theta|v[1],\dots,v[n])$$

In the biased coin toss, we recall that $P(v[n+1]|\theta)=\theta$ and $P(\theta|v[1],\dots,v[n]) = P(\theta|D)\sim\beta(\theta;M_h+\alpha,m_t+\beta)$. Thus, the last expression follows:

$$\begin{aligned}
\int_0^1 P(v[n+1]=1|\theta)P(\theta|v[1],\dots,v[n]) &= \int_0^1 \theta\frac{\Gamma(M_h+M_t+\alpha+\beta)}{\Gamma(M_h+\alpha)\Gamma(M_t+\beta)}\theta^{M_h+\alpha-1}(1-\theta)^{M_t+\beta-1}d\theta \\
&= \frac{\Gamma(M_h+M_t+\alpha+\beta)}{\Gamma(M_h+\alpha)\Gamma(M_t+\beta)}\int_0^1 \theta^{M_h+\alpha-1}(1-\theta)^{M_t+\beta-1}d\theta \\
&= \cdots \\
&= \frac{M_h+\alpha}{M_h+M_t+\alpha+\beta}
\end{aligned}$$

The conclusion is that in this example, we only need to know the previous values to estimate the following example. All this calculations are only a proof of the expectation of a beta random variable for a particular case.

2.3. Naive Bayes

Consider the Bayesian Network in Figure 1. Consider that X_i are gaussian distributions, that is

$$P(X_i|C) = \mathcal{N}(\mu_i,\sigma_i^2)$$

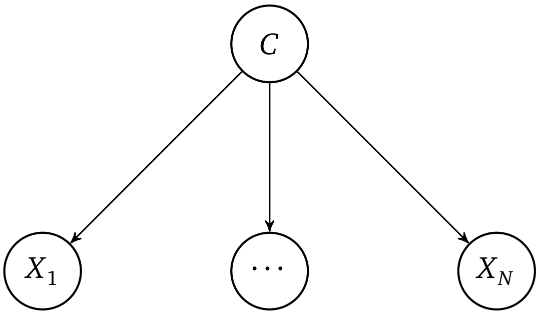

Figure 1: Naive Bayes network.

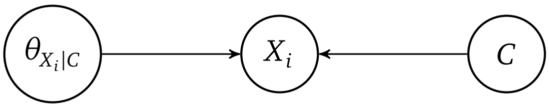

Figure 2: Relation between parameters θ, class C and X_i.

In this case, we would have the diagram presented in Figure 2 where $X_i = \{x_i[1],\ldots,x_i[n]\}$. Clearly, the parameters to seek in this case are $\theta_{X|C} = \mu_i, \sigma_i^2$. Let us compute the likelihood, assuming independence:

$$L(\mu_i, \sigma_i^2; D) = P(x_i[1],\ldots,x_i[n]|\mu_i,\sigma_i^2, C) = \prod_{j=1}^{N} \mathrm{pdf}_G(x_i;\mu_i,\sigma_i^2)$$

where pdf_G stands for the probability density function of a Gaussian distribution with parameters μ_i, σ_i^2. Hence, the log-likelihood is

$$\begin{aligned}
\ell(\mu_i,\sigma_i^2;D) &= \log\big(L(\mu_i,\sigma_i^2,D)\big) \\
&= \log\left(\prod_{j=1}^{N} \mathrm{pdf}_G(x_i;\mu_i,\sigma_i^2)\right) \\
&= \sum_i^N \log \mathrm{pdf}_G(x_i;\mu_i,\sigma_i^2) \\
&= \sum_i^N \left(\log \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i[j]-\mu_i}{\sigma_i}\right)^2}\right) \\
&= \sum_i^N \log\left(\frac{1}{\sigma_i\sqrt{2\pi}}\right) - \frac{1}{2}\sum_i^N \left(\frac{x_i[j]-\mu_i}{\sigma}\right)^2 \\
&= -N\log\sqrt{2\pi\sigma_i^2} - \frac{1}{2}\sum_i^N \left(\frac{x_i[j]-\mu_i}{\sigma}\right)^2
\end{aligned}$$

We can now compute the derivate of the log-likelihood in order to find its theoretical maximum:

$$\begin{aligned}
0 &= \frac{\partial \ell}{\partial \sigma_i^2} \\
&= -N \frac{(2\pi)/2}{\left(\sqrt{2\pi\sigma_i^2}\right)^2} + \sum_{j=1}^{N} \frac{2}{(x_i[j]-\mu_i)}(2\sigma_i^2)^2 \\
&= -N \frac{1}{2\sigma_i^2} + 2\frac{2}{(2\sigma_i^2)^2} \sum_{j=1}^{N} (x_i[j]-\mu_i)^2
\end{aligned}$$

which implies

$$\sigma_i^2 = \frac{1}{N} \sum_{j}^{N} (x_i[j]-\mu_i)^2$$

which is the expression of the sample variance. Applying this process to the expectation μ_i, we obtain that the MLE estimator is the sample mean.

Example 2.1. Gaussian Linear Model. Consider the Bayesian network in Figure 3, where $X_1, \dots, X_K$ are continuous random variables and the parents of a continuous random variable Y. We would like to model

$$P(Y|X_1, \dots, X_N) = \mathcal{N}(, \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)\, \sigma^2.$$

We can use a Linear Gaussian Model (LGM).

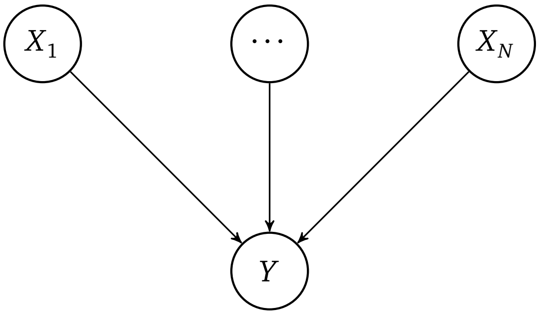

Figure 3: $X_1, \dots, X_N$ are parents of Y.

$$\begin{aligned}
P(Y|X_1, \dots, X_K) &\sim \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k, \sigma^2\right) \\
&= \mathcal{N}(\beta^T \mathbf{x}^T, \sigma^2)
\end{aligned}$$

End of example.

Example 2.2. Moving object . Consider a moving object. Let x_t be the position at time t, v_t be the speed at time t. Then, it is known that:

$$x_{t+1} = x_t + v_t \Delta t = x_t + v_t$$

then, using a LGM to obtain:

$$P\left(X_{t+1}|x_t,v_t\right)\sim\mathcal{N}\left(,x_t+v_t+1,\sigma_t^2\right).$$

End of example.

Let us now calculate the maximum likelihood estimator for the general LGM. We already know that the log likelihood for this model is:

$$\ell(\mu_i,\sigma_i^2;D)=-N\log\sqrt{2\pi\sigma_i}^2-\frac{1}{2}\sum_i^N\left(\frac{x_i[j]-\mu_i}{\sigma}\right)^2.$$

We can use the expression of this particular case and derivate with respect the different variables:

$$\begin{aligned}\frac{\partial\ell}{\partial\beta_0}&=-\sum_{j=1}^N\frac{1}{\sigma^2}\left(\beta_0+\sum_i^N(\beta_ix_i[j])-y[j]\right)\\&=-\frac{1}{\sigma^2}\left(N\beta_0+\sum\left(\beta_i\sum_jx_i[j]\right)-\sum_j^Ny[j]\right).\end{aligned}$$

Our goal is to equate the last expression to zero. We can divide the whole equation by N, since this does not change the result.Clearly, since $-\frac{\sigma^2}{N}\neq0$, we obtain:

$$0=\underbrace{\frac{1}{N}\sum j=1^Ny[j]}_{\mathbb{E}_D[y]}=\beta_0+\sum_i\left(\beta_i\underbrace{\frac{1}{N}\sum_jx_i[j]}_{\mathbb{E}_D[x_i]}\right)$$

which results in:

$$\mathbb{E}_D[y]=\beta_0+\sum_i^K\mathbb{E}_D[x_i]\tag{1}$$

We now derivate with respect the rest of the β_i.

$$\begin{aligned}\frac{\partial\ell}{\partial\beta_i}&=-\sum j=1^N\frac{1}{\sigma^2}\left(\beta_0+\sum_i^K(\beta_ix_i[j])-y[j]\right)x_i[j]\\&=-\frac{1}{\sigma^2}\left(\beta_0\sum_{j=1}^Nx_i[j]+\beta_i\sum_{j=1}^Nx_1[j]x_i[j]+\cdots+\beta_k\sum_{j=1}^Nx_k[j]x_i[j]-\sum_{j=1}^Ny[j]x_i[j]\right).\end{aligned}$$

And, as we always do in MLE, we find the maximum using the zeros of the derivative: we know that $\sigma^2/N\neq0$ so:

$$0=\mathbb{E}[yX_i]=\beta_0\mathbb{E}_D[X_i]+\sum_{i=j}^K\beta_i\mathbb{E}_D[X_j\,X_i]\tag{2}$$

3. Expectation Maximization

There are situations where we cannot apply maximum likelihood estimation as we have done before, for example, when there is missing data or latent (unobserved) variables. We can consider three types of missing data:

1. Missing completely at random: When the reason why those values are missing is independent of the values themselves and the observed ones.
2. Missing at random: The fact that data is missing is not completely random but can be explained given the observed data.
3. Missing not at random: The reason why data is missing is related with such data.

Let us see an example where the MLE algorithm cannot be applied with missing data:

Example 3.1. Let X, Y be two random variables such that

$$P(x, y \mid \Theta) = P(x \mid \theta_x) P(y \mid x, \theta_{y|x}) = \theta_x \ \theta_{y|x}.$$

That is, we are considering a simple Bayesian network $X \to Y$, with both variables being bernoulli trials. Consider the following set of observations $\mathscr{D} = \{(?, y_0)), (x_0, y_1), (?, y_0)$. The likelihood is

$$\begin{aligned} \mathscr{L}(\Theta, \mathscr{D}) &= \prod_{i=1}^{N} P(X[i], Y[i] | \Theta) \\ &= \prod_{i=1}^{N} P(X[i] | \theta_x) P(Y[i] \mid X[i], \theta_{y|x}) \\ &= P(y_1 \mid x_0) P(x_0) \left(\sum_x P(y_0 \mid x) P(x) \right)^2 \\ &= (\theta_{x_0} \theta_{y_0|x_0} + \theta_{x_1} \theta_{y_0|x_1})^2 \theta_{x_0} \theta_{y_1|x_0}. \end{aligned}$$

Where its partial derivatives cannot be independently optimized. *End of example.*

The **Expectation Maximization (EM)** algorithm focuses on the case of the **missing at random** data, and performs a MLE estimation. We will follow a two step iterative process, summarized as:

1. Compute the *expected value of the missing data*.
2. *Optimize* the set of parameters.

3.1. General EM algorithm

Given a set of observed variables $\mathbf{X} = (X_1, \dots, X_N)$ and a set of hidden or latent variables $\mathbf{Z} = (Z_1, \dots, Z_M)$, governed by a set of parameters θ, the EM algorithm seeks to find the maximum likelihood estimate of the marginal likelihood $P(\mathbf{x} \mid \theta)$ of the visible variables by applying the following 2-step iterative procedure:

1. **Expectation step**: Define $Q(\theta \mid \theta^{(t)})$ as the expected value of the log likelihood with respect to the conditional distribution of the hidden variables given the observed:

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta^{(t)}} [\log P(\mathbf{x}, \mathbf{Z} \mid \theta)] .$$

2. **Maximization step**: Find the optimal parameters that maximize Q:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t)}).$$

> **Theorem 3.1.** The marginal likelihood cannot decrease after any iteration of the expectation maximization algorithm.

Proof. For any unknown but fixed value of the hidden variables $\mathbf{z}$, applying the definition of conditional probability to $P(\mathbf{z} \mid \mathbf{x}, \theta)$ we can write[1]

$$\log P(\mathbf{x} \mid \theta) = \log P(\mathbf{x}, \mathbf{z} \mid \theta) - \log P(\mathbf{z} \mid \mathbf{x}, \theta)$$

By taking expectations over $\mathbf{Z} \mid \mathbf{x}, \theta^{(t)}$ and since $P(\mathbf{x} \mid \theta)$ does not depend on $\mathbf{Z}$, we get that

$$\begin{aligned} \log P(\mathbf{x} \mid \theta) &= \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta^{(t)}} [\log P(\mathbf{x}, \mathbf{Z} \mid \theta)] - \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta^{(t)}} [\log P(\mathbf{Z} \mid \mathbf{x}, \theta)] \\ &= Q(\theta, \theta^{(t)}) - \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta^{(t)}} [\log P(\mathbf{Z} \mid \mathbf{x}, \theta)] \end{aligned}$$

Given this quality, the increase in the marginal likelihood is

$$\begin{aligned} \log P(\mathbf{x} \mid \theta) - \log P(\mathbf{x} \mid \theta^{(t)}) &= Q(\theta, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)}) \\ &\quad - \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta^{(t)}} [\log P(\mathbf{Z} \mid \mathbf{x}, \theta)] + \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta^{(t)}} \left[\log P(\mathbf{Z} \mid \mathbf{x}, \theta^{(t)})\right] \\ &= Q(\theta, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)}) + KL\left(P(\mathbf{Z} \mid \mathbf{x}, \theta^{(t)}) \middle| P(\mathbf{Z} \mid \mathbf{x}, \theta)\right) \end{aligned}$$

Using that the KL divergence is always positive, we arrive at

$$\log P(\mathbf{x} \mid \theta^{(t+1)}) - \log P(\mathbf{x} \mid \theta^{(t)}) \geq Q(\theta^{(t+1)}, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)}) \geq 0.$$

Where the last inequality is given by the maximization step of the EM algorithm.
■

[1] Given that $P(\mathbf{z} \mid \mathbf{x}, \theta) \neq 0$.