Introduction to Big Data and Machine Learning
Bayesian Probabilities

Dr. Mihail

September 10, 2019

Probability review

- Sample space Ω: set of all possible outcomes for an experiment
- Event space $\mathcal{A}$: space of potential results of the experiment. A subset A of Ω is in the event space $\mathcal{A}$ if at the end of the experiment, we can observe whether a particular event $\omega \in \Omega$
- Probability of P: With each event $A \in \mathcal{A}$, we associate a number $P(A)$ that measures the probability or degree of belief that the event will occur. $P(A)$ is called the probability of A.
- The probability of a single event must be in the interval $[0, 1]$, and the total probability over all outcomes in Ω must be 1, i.e.: $P(\Omega) = 1$

Conditional Probabilities

Consider data D and model parameters w

$$P(w, D) = P(w|D)P(D) \tag{1}$$

$$P(D, w) = P(D|w)P(w) \tag{2}$$

therefore

$$P(w|D)P(D) = P(D|w)P(w) \tag{3}$$

hence

$$P(w|D) = \frac{P(D|w)P(w)}{P(w)} \tag{4}$$

Bayesian linear regression

$$P(w|D) = \frac{P(D|w)P(w)}{P(w)} \tag{5}$$

- $P(w|D)$ is referred to as the posterior
- $P(w)$ prior probability, our prior belief about the model parameters
- $P(D|w)$ is the likelihood function

also,

$$P(D) = \int P(D|w)P(w)dw \tag{6}$$

Bayesian linear regression

Bayesian v. Frequentist

- Frequentist setting: w is considered fixed, obtained by an "estimator", whose error bars are obtained by considering the distribution of data sets D
- Bayesian approach: there is a single dataset D, the one observed, and the uncertainty in parameters is expressed through a probability distribution over w

A widely used approach in frequentist approach is to estimate the maximum likelihood, in which w is computed that maximizes the likelihood function $P(D|w)$

Linear basis function models

Linear regression models share the property of being linear in their parameters but not necessarily in their input variables. Using non-linear basis functions of input variables, linear models are able model arbitrary non-linearities from input variables to targets. A linear regression model y(x,w) can therefore be defined more generally as:

$$y(x, w) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x) = \sum_{j=0}^{M-1} w_j \phi_j(x) = w^T \phi(x) \tag{7}$$

where ϕ_j are the basis functions and M is the total number of parameters w_j including the bias term w_0.

- $\phi_0(x) = 1$

and in the case of simple linear regression $\phi(x) = x$

Bayesian linear regression

- The target variable t of an observation x is given by a deterministic function $y(x, w)$
$$t = y(x, w) + \epsilon \tag{8}$$
where ϵ is additive noise, normally distributed (i.e., follows a Gaussian distribution with zero mean and precision[inverse variance] β)
The probabilistic model of t given x can be written as:
$$p(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} exp(-\frac{\beta}{2}(t - y(x, w)^2) \tag{9}$$

Bayesian linear regression

Likelihood function

To fit a model, we use N independent and identically distributed observations $x_1, x_2, \ldots, x_N$ and their corresponding targets $t_1, t_2, \ldots, t_N$, combined in a matrix X where $X_{(i,:)} = x_i^T$ and scalar targets t_i into column vector t, the joint conditional distribution of targets t given X (the likelihood function) is:

$$P(t|X, w, \beta) = \Pi_{i=1}^{N} \mathcal{N}(t_i|w^T \phi(x_i), \beta^{-1}) \tag{10}$$

Taking the log of the likelihood, we get:

$$log P(t|w, \beta) = \frac{N}{2} log \beta - \frac{N}{2} log 2\pi - \beta E_D(w) \tag{11}$$

where $E_D(w)$ is the sum of squares error function coming from the exponent of the likelihood function.

Bayesian linear regression

$$E_D(w) = \frac{1}{2}\sum_{i=1}^{N}(t_i - w^T\phi(x_i))^2 = \frac{1}{2}||t - \Phi w||^2 \tag{12}$$

where Φ is the design matrix defined as

$$\Phi = \begin{bmatrix} \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \dots & \phi_{M-1}(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \dots & \phi_{M-1}(x_N) \end{bmatrix} \tag{13}$$

Bayesian linear regression

Bayesian approach

For a Bayesian treatment, we need a prior probability distribution over w. For simplicity, we will use an isotropic Gaussian distribution over w with zero mean:

$$P(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I) \tag{14}$$

The posterior can be written as:

$$P(w|t, \alpha, \beta) = \mathcal{N}(w|m_N, S_N) \tag{15}$$

where

$$m_N = \beta S_N \Phi^T t \tag{16}$$

$$S_N^{-1} = \alpha I + \beta \Phi^T \Phi \tag{17}$$

can be analytically derived (skipped here) because the conjugate are also Gaussian.

Bayesian linear regression

Bayesian approach

Taking the log:

$$logP(W|t, \alpha, \beta) = \beta E_D(w) - \alpha E_w(w) + const \tag{18}$$

where $E_D(w)$ comes from Eq 12 and

$$E_W(w) = \frac{1}{2} w^T w \tag{19}$$

Bayesian linear regression

Posterior distribution

To make a prediction t at a new location x, we use the posterior:

$$p(t|x, \mathbf{t}, \alpha, \beta) = \int p(t|x, w, \beta) p(w|\mathbf{t}, \alpha, \beta) dw \tag{20}$$

hence we not only get an estimate, but also an uncertainty:

$$p(t|x, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|m_N^T \phi(x), \sigma_N^2(x)) \tag{21}$$

where $m_N^T \phi(x)$ is the regression function after N observations and $\sigma_N^2(x)$ is the corresponding predictive variance:

$$\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x) \tag{22}$$