Machine Learning
Artur d'Avila Garcez
3rd Tutorial – Bayesian Inference

Background

Bayesian inference is a method of statistical inference in which Bayes' rule is used to update the probability of a hypothesis as new evidence is acquired. Bayesian inference is an important technique in statistics, particularly important in the dynamic analysis of a sequence of data. Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, medicine, and law. In the philosophy of decision theory, Bayesian inference is closely related to subjective probability often called *Bayesian probability*. Bayesian probability provides a rational method for updating beliefs, as follows.

Definitions

- x, a data point in general. This may in fact be a vector of values.
- θ, the parameter of the data point's distribution, i.e. $x \sim p(x \mid \theta)$. This may be a vector of parameters.
- α, the parameter of the parameter (a.k.a. hyperparameter), i.e. $\theta \sim p(\theta \mid \alpha)$. This may be a vector of hyperparameters.
- $\mathbf{X}$, a set of n observed data points, i.e. $x_1, \ldots, x_n$.
- $\tilde{x}$, a new data point whose distribution is to be predicted.

Bayesian inference

- The prior distribution is the distribution of the parameter(s) before any data is observed, i.e. $p(\theta \mid \alpha)$.
- The prior distribution might not be easily determined. In this case, Jeffrey's prior is used in general (see http://en.wikipedia.org/wiki/Jeffreys_prior and http://en.wikipedia.org/wiki/Fisher_information).
- The sampling distribution is the distribution of the observed data conditioned on its parameters, i.e. $p(\mathbf{X} \mid \theta)$. This is also termed the likelihood, especially when viewed as a function of the parameters, sometimes written $\mathrm{L}(\theta \mid \mathbf{X}) = p(\mathbf{X} \mid \theta)$.
- The marginal likelihood (sometimes also termed *the evidence*) is the distribution of the observed data marginalized over the parameters, i.e.:

$$p(\mathbf{X} \mid \alpha) = \int_{\theta} p(\mathbf{X} \mid \theta) p(\theta \mid \alpha) \, \mathrm{d}\theta$$

- The posterior distribution is the distribution of the parameters after taking into account the observed data. This is determined by Bayes' rule, which forms the heart of Bayesian inference:

$$p(\theta \mid \mathbf{X}, \alpha) = \frac{p(\mathbf{X} \mid \theta) p(\theta \mid \alpha)}{p(\mathbf{X} \mid \alpha)} \propto p(\mathbf{X} \mid \theta) p(\theta \mid \alpha)$$

This is frequently expressed in words as "the posterior is proportional to the likelihood multiplied by the prior", or "posterior equals likelihood times prior, over evidence".

Setup – Matlab

We will run a simple example of Bayesian inference over a prior normal distribution. Samples representing another normal distribution will be used as likelihood, iteratively, and the resulting distribution will be shown.

1. Unzip “bayesExample.zip” into any folder of your choice;
2. Open Matlab and inside it, locate the folder where the contents of “bayesExample.zip” have been unzipped;
3. Double click on “bayesExample.m” inside Matlab;
4. Left-click the editor and press F5 to run the code

You should see a sequence of plots of priors and likelihoods, where the first prior will slowly become closer and closer to the likelihood.

The main thing to note as the process iterates (prior(t+1) = posterior(t); we use ‘=’ to denote the assignment of the posterior at time t as the prior at time t+1) is that without changing the data, the posterior will have a large bias towards the maximum likelihood with a much smaller variance over time.

Exercises

1. Change the variance/mean of the distributions:

    a. Prior;
    b. Likelihood;

    to see how the updates will change as a result.

To change the prior, these are the lines of code that require changing:

line 9 pm = 5; % Prior mean

line 10 ps2 = 4; % Prior variance

line 11 prior = (2*pi*ps2)^(-0.5)*exp(-0.5*(x-pm).^2/ps2);

To change the likelihood, change the following lines of code:

line 18 lm = 6; % sample mean

line 19 ls2 = 3; % sample variance

line 23 like = (2*pi*ls2)^(-0.5)*exp(-0.5*(x-lm).^2/ls2);

But notice that any changes above will require a new calculation of the posterior, defined in the original code as a multiplication of Gaussians, as follows:

line 27 Ps2 = (1/ps2 + 1/ls2)^(-1);

line 28 Pm = Ps2*(pm/ps2 + lm/ls2);

line 29 posterior = (2*pi*Ps2)^(-0.5)*exp(-0.5*(x-Pm).^2/Ps2);

The above code multiplies two Gaussians (prior and likelihood) by calculating mean and variance of the resulting Gaussian (posterior). For some distributions (e.g. Gaussian, Beta), it is known that the posterior should have the same form as the prior following multiplication by the likelihood function. But depending on how you define the prior, the multiplication may not be trivial.

Challenge: change the code to use a uniform distribution as prior and a linear function as likelihood, as exemplified in class with the coin tossing example. Can you simulate the coin tossing example whereby a normal posterior with very small variance is produced after 100 trials? You will need to compute a pointwise multiplication of the 1,000 points used in the code to plot the curves (these are currently uniformly distributed along the x axis in the interval from 0 to 10). The posterior will have to be normalised after each trial.