LECTURE 4
A Bayesian approach to linear regression

The main motivations behind a Bayesian formalism for inference are a coherent approach to modeling uncertainty as well as an axiomatic framework for inference. We will reformulate multivariate linear regression from a Bayesian formulation in this section.

Bayesian inference involves thinking in terms of probability distributions and conditional distributions. One important idea is that of a conjugate prior. Another tool we will use extensively in this class is the multivariate normal distribution and its properties.

4.1. Conjugate priors

Given a likelihood function $p(x \mid \theta)$ and a prior $\pi(\theta)$ on can write the posterior as

$$p(\theta \mid x) = \frac{p(x \mid \theta)\pi(\theta)}{\int_{\theta'} p(x \mid \theta')\pi(\theta')\,\mathrm{d}\theta'} = \frac{p(x,\theta)}{p(x)},$$

where $p(x)$ is the marginal density for the data, $p(x,\theta)$ is the joint density of the data and the parameter θ.

The idea of a prior and likelihood being conjugate is that the prior and the posterior densities belong to the same family. We now state some examples to illustrate this idea.

Beta, Binomial: Consider the Binomial likelihood with n (the number of trials) fixed

$$f(x \mid p, n) = \binom{n}{x} p^x (1-p)^{n-x},$$

the parameter of interest (the probability of a success) is $p \in [0,1]$. A natural prior distribution for p is the Beta distribution which has density

$$\pi(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad p \in (0,1) \text{ and } \alpha,\beta > 0,$$

where $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ is a normalization constant. Given the prior and the likelihood densities the posterior density modulo normalizing constants will take the form

$$\begin{aligned} f(p \mid x) &\propto \left[\binom{n}{x} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right] p^x (1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1}, \\ &\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1}, \end{aligned}$$

which means that the posterior distribution of p is also a Beta with

$$p \mid x \sim \text{Beta}(\alpha + x, \beta + n - x).$$

Normal, Normal: Given a normal distribution with unknown mean the density for the likelihood is

$$f(x \mid \theta, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right),$$

and one can specify a normal prior

$$\pi(\theta; \theta_0, \tau_0^2) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\theta_0)^2\right),$$

with hyper-parameters θ_0 and τ_0. The resulting posterior distribution will have the following density function

$$f(\theta \mid x) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right) \times \exp\left(-\frac{1}{2\tau_0^2}(\theta-\theta_0)^2\right),$$

which after completing squares and reordering can be written as

$$\theta \mid x \sim \text{N}(\theta_1, \tau_1^2), \quad \theta_1 = \frac{\frac{\theta_0}{\tau_0^2} + \frac{x}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \quad \tau_1^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}.$$

4.2. Bayesian linear regression

We start with the likelihood as

$$f(Y \mid \mathbf{X}, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2}\right).$$

and the prior as

$$\pi(\beta) \propto \exp\left(-\frac{1}{2\tau_0^2}\beta^T\beta\right).$$

The density of the posterior is

$$\text{Post}(\beta \mid D) \propto \left[\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2}\right)\right] \times \frac{1}{(2\pi)^{p/2}\gamma^{1/2}} \exp\left(-\frac{1}{2\tau_0^2}\beta^T\beta\right).$$

With a good bit of manipulation the above can be rewritten as a multivariate normal distribution

$$\beta \mid Y, \mathbf{X}, \sigma^2 \sim \text{N}_p(\mu_1, \Sigma_1)$$

with

$$\Sigma_1 = (\tau_0^{-2}\mathbf{I}_p + \sigma^{-2}\mathbf{X}^T\mathbf{X})^{-1}, \quad \mu_1 = \sigma^{-2}\Sigma_1\ \mathbf{X}^T Y.$$

Note the similarities of the above distribution to the MAP estimator. Relate the mean of the above estimator to the MAP estimator.

Predictive distribution: Given data $D = \{(x_i, y_i)\}_{i=1^n}$ and and a new value x_* one would like to estimate y_*. This can be done using the posterior and is called the posterior predictive distribution

$$f(y_* \mid D, x_*, \sigma^2, \tau_0^2) = \int_{\mathbb{R}^p} f(y_* \mid x_*, \beta, \sigma^2) f(\beta \mid Y, \mathbf{X}, \sigma^2, \tau_0^2)\ \mathrm{d}\beta,$$

where with some manipulation

$$y_* \mid D, x_*, \sigma^2, \tau_0^2 \sim \mathrm{N}(\mu_*, \sigma_*^2),$$

where

$$\mu_* = \frac{1}{\sigma^2} \Sigma_1 \mathbf{X}^T Y x_*, \quad \sigma_*^2 = \sigma^2 + x_*^T \Sigma_1 x_*.$$