Lecture 8 - *Maximum Likelihood and Bayesian Learning*

OBJECTIVE: In this chapter, we revise maximum likelihood (ML) for a simple binary model. We then introduce Bayesian learning for this simple model and for the linear-Gaussian regression setting of the previous chapters. The key difference between the two approaches is that the frequentist view assumes there is one true model responsible for the observations, while the Bayesian view assumes that the model is a random variable with a certain prior distribution. Computationally, the ML problem is one of optimization, while Bayesian learning is one of integration.

◇ MAXIMUM LIKELIHOOD

Frequentist Learning assumes that there is a true model (say a parametric model with parameters θ_0). The estimate is denoted $\widehat{\theta}$. It can be found by maximising the **likelihood**:

⋆

$$\widehat{\theta} = \arg\max_{\theta} \quad p(x_{1:n}|\theta)$$

For **identical and independent distributed** (i.i.d.) data:

$$p(x_{1:n}|\theta) =$$

$$\mathcal{L}(\theta) = \log p(x_{1:n}|\theta) =$$

Let's illustrate this with a coin-tossing example.

⋆ Let $x_{1:n}$, with $x_i \in \{0, 1\}$, be i.i.d. Bernoulli:

$$p(x_{1:n}|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

With $m \triangleq \sum x_i$, we have

$$\mathcal{L}(\theta) =$$

Differentiating, we get

◇ BAYESIAN LEARNING

Given our **prior** knowledge $p(\theta)$ and the data **model** $p(\cdot|\theta)$, the Bayesian approach allows us to update our prior using the new data $x_{1:n}$ as follows:

$$p(\theta|x_{1:n}) = \frac{p(x_{1:n}|\theta)p(\theta)}{p(x_{1:n})}$$

where $p(\theta|x_{1:n})$ is the **posterior distribution**, $p(x_{1:n}|\theta)$ is the likelihood and $p(x_{1:n})$ is the **marginal likelihood** (evidence). Note

$$p(x_{1:n}) = \int p(x_{1:n}|\theta)p(\theta)d\theta$$

⋆

Bayesian Prediction

We predict by marginalising over the posterior of the parameters

$$\begin{aligned} p(x_{n+1}|x_{1:n}) &= \int p(x_{n+1}, \theta|x_{1:n}) d\theta \\ &= \int p(x_{n+1}|\theta) p(\theta|x_{1:n}) d\theta \end{aligned}$$

Bayesian Model Selection

For a particular model structure M_i, we have

$$p(\theta|x_{1:n}, M_i) = \frac{p(x_{1:n}|\theta, M_i) p(\theta|M_i)}{p(x_{1:n}|M_i)}$$

Models are selected according to their posterior:

$$P(M_i|x_{1:n}) \propto P(x_{1:n}|M_i) p(M_i) = P(M_i) \int p(x_{1:n}|\theta, M_i) p(\theta|M_i) d\theta$$

The ratio $P(x_{1:n}|M_i)/P(x_{1:n}|M_j)$ is known as the **Bayes Factor**.

$\star$ Let $x_{1:n}$, with $x_i \in \{0, 1\}$, be i.i.d. Bernoulli: $x_i \sim \mathcal{B}(1, \theta)$

$$p(x_{1:n}|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \theta^m (1-\theta)^{n-m}$$

Let us choose the following **Beta** prior distribution:

$$p(\theta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

where Γ denotes the Gamma-function. For the time being, α and β are fixed **hyper-parameters**. The posterior distribution is proportional to:

$$p(\theta|x) \propto$$

with normalisation constant

Since the posterior is also Beta, we say that the Beta prior is **conjugate** with respect to the binomial likelihood. Conjugate priors lead to the same form of posterior.

Different hyper-parameters of the Beta $\mathcal{B}e(\alpha, \beta)$ distribution give rise to different prior specifications:

The generalisation of the Beta distribution is the Dirichlet distribution $\mathcal{D}(\alpha_i)$, with density

$$p(\theta) \propto \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

where we have assumed k possible thetas. **Note that the Dirichlet distribution is conjugate with respect to a Multinomial likelihood.**

◇ BAYESIAN LEARNING FOR LINEAR-GAUSSIAN MODELS

In the Bayesian linear prediction setting, we focus on computing the posterior:

$$\begin{aligned} p(\theta|X,Y) &\propto p(Y|X,\theta)p(\theta) \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(Y-X\theta)^T(Y-X\theta)} p(\theta) \end{aligned}$$

We often want to maximise the posterior — that is, we look for the *maximum a poteriori* (MAP) estimate. In this case, the choice of prior determines a type of constraint! For example, consider a Gaussian prior $\theta \sim \mathcal{N}(0, \delta^2\sigma^2 I_d)$. Then

$$p(\theta|X,Y) \propto \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(Y-X\theta)^T(Y-X\theta)} \left(2\pi\sigma^2\delta^2\right)^{-\frac{d}{2}} e^{-\frac{1}{2\delta^2\sigma^2}\theta^T\theta}$$

Our task is to rearrange terms in the exponents in order to obtain a simple expression for the posterior distribution.

⋆

$$\begin{aligned} p(\theta|X,Y) &= \left|2\pi\sigma^2 M\right|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(\theta-\mu)^T M^{-1}(\theta-\mu)} \\ &\propto \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(Y-X\theta)^T(Y-X\theta)} \left(2\pi\sigma^2\delta^2\right)^{-\frac{d}{2}} e^{-\frac{1}{2\delta^2\sigma^2}\theta^T\theta} \end{aligned}$$

So the posterior for θ is Gaussian:

$$p(\theta|X,Y) \;=\; \left|2\pi\sigma^2 M\right|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(\theta-\mu)^T M^{-1}(\theta-\mu)}$$

with **sufficient statistics**:

$$\begin{aligned}\mathbb{E}(\theta|X,Y) &= (XX^T+\delta^{-2}I_d)^{-1}X^TY \\ var(\theta|X,Y) &= (XX^T+\delta^{-2}I_d)^{-1}\sigma^2\end{aligned}$$

The MAP point estimate is:

$$\widehat{\theta}_{MAP} \;=\; (XX^T+\delta^{-2}I_d)^{-1}X^TY$$

It is the same as the ridge estimate (except for a trivial negative sign in the exponent of δ), which results from the L_2 constraint. A flat ("vague") prior with large variance (large δ) leads to the ML estimate.

$$\widehat{\theta}_{MAP} = \widehat{\theta}_{ridge} \quad \stackrel{\delta^2\to 0}{\longrightarrow} \quad \widehat{\theta}_{ML} = \widehat{\theta}_{SVD} = \widehat{\theta}_{LS}$$

2 Full Bayesian Model

In Bayesian inference, we're interested in the full posterior:

$$p(\theta,\sigma^2,\delta^2|X,Y) \propto p(Y|\theta,\sigma^2,X)p(\theta|\sigma^2,\delta^2)p(\sigma^2)p(\delta^2)$$

where

$$\begin{aligned}Y|\theta,\sigma^2,X &\sim \mathcal{N}(X\theta,\sigma^2 I_n) \\ \theta &\sim \mathcal{N}\left(0,(\sigma^2\delta^2 I_d)\right) \\ \sigma^2 &\sim \mathcal{IG}\left(a/2,b/2\right) \\ \delta^2 &\sim \mathcal{IG}\left(\alpha,\beta\right)\end{aligned}$$

where $\mathcal{IG}\left(\alpha,\beta\right)$ denotes the **Inverse-Gamma distribution**.

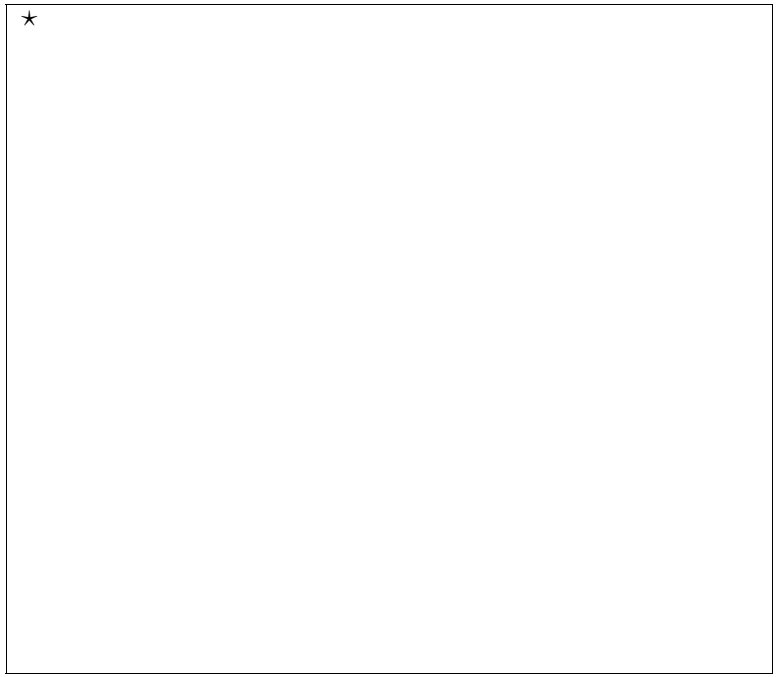
$$\delta^2 \sim \mathcal{IG}\left(\alpha,\beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta/\delta^2}(\delta^2)^{-\alpha-1}\mathbb{I}_{[0,\infty)}(\delta^2)$$

This is the conjugate prior for the variance of a Gaussian. The generalization of the Gamma distribution, i.e. the conjugate prior of a covariance matrix is the **inverse**

Wishart distribution $\Sigma \sim IW_d(\alpha, \alpha\Sigma^*)$, admitting the density

$$p(\Sigma|\alpha, \Sigma^*) \propto |\Sigma|^{-(\alpha+d+1)/2} \exp\{-(1/2)tr(\alpha\Sigma^*\Sigma^{-1})\}$$

We can visualise our hierarchical model with the following graphical model:



The product of likelihood and priors is:

$$\begin{aligned} p(\theta, \sigma^2, \delta^2|X, Y) \;\propto\; & \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(Y-X\theta)^T(Y-X\theta)} \\ & \times \left(2\pi\sigma^2\delta^2\right)^{-\frac{d}{2}} e^{-\frac{1}{2\delta^2\sigma^2}\theta^T\theta} \\ & \times (\sigma^2)^{-a/2-1} e^{-\frac{b}{2\sigma^2}} (\delta^2)^{-\alpha-1} e^{-\frac{\beta}{\delta^2}} \end{aligned}$$

We know from our previous work on computing the posterior for θ that:

$$\begin{aligned} p(\theta, \sigma^2, \delta^2|X, Y) \;\propto\; & (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}Y^TPY} \\ & \times (2\pi\sigma^2\delta^2)^{-d/2} e^{-\frac{1}{2\sigma^2}(\theta-\mu)^TM^{-1}(\theta-\mu)} \\ & \times (\sigma^2)^{-a/2-1} e^{-\frac{b}{2\sigma^2}} (\delta^2)^{-\alpha-1} e^{-\frac{\beta}{\delta^2}} \end{aligned}$$

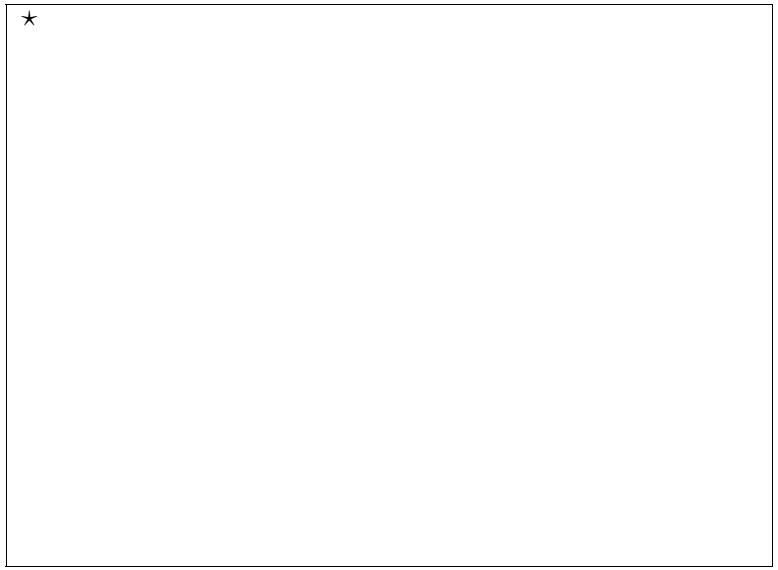
where

$$\begin{aligned} M^{-1} &= X^TX + \delta^{-2}I_d \\ \mu &= MX^TY \\ P &= I_n - XMX^T \end{aligned}$$

From this expression, it is now obvious that

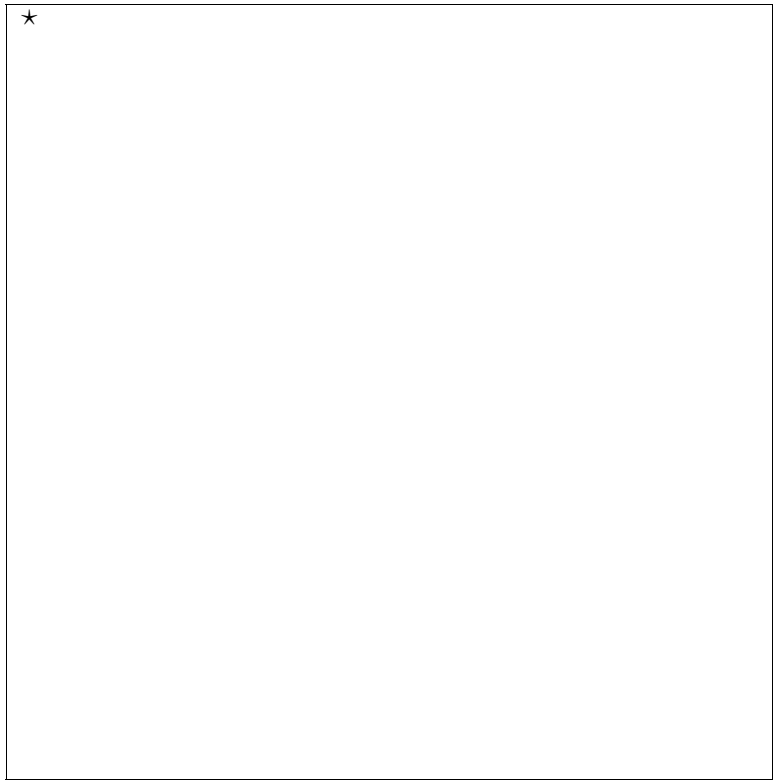
$$p(\theta|\sigma^2, X, Y) = \mathcal{N}(\mu, \sigma^2 M)$$

Next, we integrate $p(\theta, \sigma^2, \delta^2|X, Y)$ over θ in order to get an expression for $p(\sigma^2, \delta^2|X, Y)$. This will allow us to get an expression for the marginal posterior $p(\sigma^2|X, Y)$.



$$p(\sigma^2|X, Y) \sim \mathcal{IG}\left(\frac{a+n}{2}, \frac{b+Y'PY}{2}\right)$$

Integrating over σ^2 gives us an expression for $p(\delta^2|X, Y)$



Unfortunately this is a nonstandard distribution, thus making it hard for us to come up with the normalizing constant. So we'll make use of the fact that we know θ and σ^2 to derive

a conditional distribution $p(\delta^2|\theta, \sigma^2, X, Y)$.

⋆ We know that:

$$p(\theta, \sigma^2, \delta^2|X, Y) \propto \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(Y-X\theta)^T(Y-X\theta)} \times \left(2\pi\sigma^2\delta^2\right)^{-\frac{d}{2}} e^{-\frac{1}{2\delta^2\sigma^2}\theta^T\theta} \times (\sigma^2)^{-a/2-1} e^{-\frac{b}{2\sigma^2}} (\delta^2)^{-\alpha-1} e^{-\frac{\beta}{\delta^2}}$$

In summary, we can

- Obtain $p(\theta|\sigma^2, \delta^2, X, Y)$ analytically.
- Obtain $p(\sigma^2|\delta^2, X, Y)$ analytically.
- Derive an expression for $p(\delta^2|\theta, \sigma^2, X, Y)$.

Given δ^2, we can obtain analytical expressions for θ and σ^2. But, let's be a bit more ambitious. Imagine we could run the following sampling algorithm (known as the **Gibbs Sampler**)

⋆

1. LOAD data (X, Y).
2. Compute $X^T Y$ and $X^T X$.
3. Set, e.g., $a = b = 0$, $\alpha = 2$ and $\beta = 10$.
4. Sample $\delta^{2(0)} \sim \mathcal{IG}(\alpha, \beta)$.
5. FOR $i = 1$ to N:
   (a) Compute M, P and $\mu^{(i)}$ using $\delta^{2(i-1)}$.
   (b) Sample $\sigma^{2(i)} \sim \mathcal{IG}\left(\frac{a+n}{2}, \frac{b+Y'PY}{2}\right)$.
   (c) Sample $\theta^{(i)} \sim \mathcal{N}(\mu^{(i)}, \sigma^{2(i)} M)$.
   (d) Sample $\delta^{2(i)} \sim \mathcal{IG}\left(\frac{d}{2} + \alpha, \beta + \frac{\theta^{(i)T}\theta^{(i)}}{2\sigma^{2(i)}}\right)$.

We can use these samples in order to approximate the integrals of interest with Monte Carlo averages.

For example, the predictive distribution

$$p(y_{n+1}|X_{1:n+1}, Y) = \int p(y_{n+1}|\theta, \sigma^2, x_{n+1}) p(\theta, \sigma^2, \delta^2|X, Y) d\theta d\sigma^2 d\delta^2$$

can be approximated with:

$$\widehat{p}(y_{n+1}|X_{1:n+1}, Y) = \frac{1}{N} \sum_{i=1}^{N} p(y_{n+1}|\theta^{(i)}, \sigma^{2(i)}, x_{n+1})$$

That is,

⋆

$\widehat{p}(y_{n+1}|X_{1:n+1}, Y) =$

In the next lecture, we will derive the theory that justifies the use of this algorithm as well a many other **Monte Carlo** algorithms.