1 Maximum Likelihood Review

Suppose you are collecting data on the relative rates of different types of twins, and you obtain the following observations:

- there are m_i pairs of identical male twins and f_i pairs of identical female twins
- there are m_f pairs of fraternal male twins and f_f pairs of fraternal female twins
- there are b pairs of fraternal opposite gender twins

To model this data, we choose these distributions and parameters:

- Given that a pair of siblings are twins, they are identical with probability θ, and non-identical with probability $1-\theta$
- Given they are identical twins, the twins are both male with probability p and both female with probability $1-p$.
- Given they are twins and not identical (and thus are fraternal twins), the probability of both male twins is q^2, probability of both female twins is $(1-q)^2$ and probability of opposite gender twins is $2q(1-q)$.

(a) Write expressions for the likelihood and the log-likelihood of the data as functions of the parameters θ, p, and q for the observations $m_i,\ f_i,\ m_f,\ f_f,\ b$.

Solution: The probability of identical male twins is θp, probability of identical female twins is $\theta(1-p)$, probability of fraternal male twins is $(1-\theta)q^2$, probability of fraternal female twins is $(1-\theta)(1-q)^2$ and probability of fraternal opposite gender twins is $(1-\theta)\cdot 2q(1-q)$.

$$\begin{aligned}
L(\theta, p, q) &= (\theta p)^{m_i} \cdot (\theta(1-p))^{f_i} \cdot ((1-\theta)q^2)^{m_f} \cdot ((1-\theta)(1-q)^2)^{f_f} \\
&\quad \cdot ((1-\theta)\cdot 2q(1-q))^b \\
&= \theta^{(m_i+f_i)}(1-\theta)^{(m_f+f_f+b)} \cdot p^{m_i}(1-p)^{f_i} \cdot q^{2m_f}(1-q)^{2f_f}(2q(1-q))^b
\end{aligned}$$

$$\begin{aligned}
l(\theta, p, q) &= (m_i+f_i)\cdot\log\theta + (m_f+f_f+b)\cdot\log(1-\theta) + m_i\cdot\log p + \\
&\quad f_i\cdot\log(1-p) + 2m_f\cdot\log q + 2f_f\cdot\log(1-q) + b\cdot\log(2q(1-q))
\end{aligned}$$

Likelihood $L(\theta, p, q)$ = **Solution:** $\theta^{(m_i+f_i)}(1-\theta)^{m_f+f_f+b} \cdot p^{m_i}(1-p)^{f_i} \cdot q^{2m_f}(1-q)^{2f_f}(2q(1-q))^b$

Log likelihood $l(\theta, p, q)$ = **Solution:** $(m_i + f_i)\log\theta + (m_f + f_f + b)\log(1-\theta) + m_i \log p + f_i \log(1-p) + 2m_f \log(q) + 2f_f \log(1-q) + b\log(2q(1-q))$

(b) What are the maximum likelihood estimates for θ, p and q? Scratch space is provided to you here, which you may find useful.

Solution: To get the maximum likelihood estimate, we have to maximize the log likelihood by taking partial derivatives. The partial derivatives and corresponding maximum likelihood estimates are given by

$$\frac{\partial l}{\partial \theta} = \frac{m_i+f_i}{\theta} - \frac{m_f+f_f+b}{1-\theta} = 0 \quad \theta_{\text{ML}} = \frac{m_i+f_i}{m_i+f_i+m_f+f_f+b}$$

$$\frac{\partial l}{\partial p} = \frac{m_i}{p} - \frac{f_i}{1-p} = 0 \quad p_{\text{ML}} = \frac{m_i}{m_i+f_i}$$

$$\frac{\partial l}{\partial q} = \frac{2m_f+b}{q} - \frac{2f_f+b}{1-q} = 0 \quad q_{\text{ML}} = \frac{2m_f+b}{2m_f+2f_f+2b}$$

2 MAP Estimation Review

Suppose we have a data set of n data points $D = \{x_1, \ldots, x_n\}$, with each point drawn independently from a Gaussian with mean μ and variance σ^2.

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

We will place the following prior on μ:

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

The MAP estimate of μ is defined as

$$\mu_{\text{MAP}} = \arg\max_{\mu} p(\mu|D)$$

(a) Write an expression for the MAP estimate of μ.

Solution: Using Bayes' rule and plugging in probability densities,

$$\begin{aligned} p(\mu|D) &\propto p(\mu)p(D|\mu) \\ &= \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

Taking logs dropping irrelevant constants, and negating gives an objective to be minimized

$$l(\mu|D) = \log\sigma_0 + \frac{(\mu-\mu_0)^2}{2\sigma_0^2} + \sum_{i=1}^{n}\left(\log\sigma + \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

To minimize this, we take the derivative with respect to μ and set it to zero

$$\frac{\partial l}{\partial\mu} = \frac{\mu-\mu_0}{\sigma_0^2} - \sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2} = 0$$

Solving for μ,

$$\begin{aligned} \mu_{\text{MAP}} &= \left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_i x_i\right) \Bigg/ \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) \\ &= \alpha\mu_0 + (1-\alpha)\left(\frac{1}{n}\sum_i x_i\right) \end{aligned}$$

where

$$\alpha = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}$$

Thus, the MAP estimate of μ lies on a line between the prior mean μ_0 and the sample mean $\frac{1}{n}\sum_i x_i$.

(b) What happens to the MAP estimate as $\sigma_0^2 \to \infty$, and how is this estimate related to the ML estimate? Interpret this result.

Solution: We will have $\alpha \to 0$ as $\sigma_0^2 \to \infty$, so $\mu_{\text{MAP}} \to \frac{1}{n}\sum_i x_i = \mu_{\text{ML}}$. In other words, infinite variance on the prior of μ leads to a MAP estimate equal to the ML estimate—that is, if we have no prior knowledge of μ at all, the best we can do is the sample mean.

(c) What happens to the MAP estimate as $\sigma^2 \to \infty$?

Solution: As $\sigma^2 \to \infty$, we get $\alpha \to 1$, so $\mu_{\text{MAP}} \to \mu_0$. Informally, infinite variance on the data means that we can't trust the data at all, so our MAP completely avoids the data and estimates μ as the prior mean.

3 Prediction Error of Ridge Regression

(a) Let A be a $d \times n$ matrix and B be a $n \times d$ matrix. For any $\mu > 0$, show that $(AB + \mu I)^{-1}A = A(BA + \mu I)^{-1}$, if $AB + \mu I$ and $BA + \mu I$ are invertible.

Solution: We begin with an equality

$$ABA + \mu A = ABA + \mu A.$$

Factoring both sides,

$$A(BA + \mu I) = (AB + \mu I)A.$$

Lastly, if if $AB + \mu I$ and $BA + \mu I$ are invertible, we can left-multiply by $(BA + \mu I)^{-1}$ and right-multiply by $(AB + \mu I)^{-1}$. This yields

$$(AB + \mu I)^{-1}A(BA + \mu I)(BA + \mu I)^{-1} = (AB + \mu I)^{-1}(AB + \mu I)A(BA + \mu I)^{-1}.$$

Simplifying,

$$(AB + \mu I)^{-1}A = A(BA + \mu I)^{-1},$$

as needed.

(b) Let $X \in \mathbb{R}^{n \times d}$ be n samples of d features, and $y \in \mathbb{R}^n$ be the corresponding n samples of the quantity that you would like to predict with regression. Let

$$\widehat{\theta}_\lambda = \arg\min_\theta \|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2,$$

for $\lambda > 0$, be the solution to the ridge regression problem.

Using part (a), show that $\widehat{\theta}_\lambda = X^\top(XX^\top + \lambda I)^{-1}y$.

Solution:

Start by taking the gradient of the loss function as follows:

$$\begin{aligned}\nabla_\theta(\|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2) &= (X\theta - y)^\top X + \lambda\theta^\top \\ &= \theta^\top X^\top X - Y^\top X + \lambda\theta^\top \\ &= X^\top X\theta - X^\top y + \lambda\theta = 0\end{aligned}$$

Therefore,

$$\widehat{\theta}_\lambda = (X^\top X + \lambda I)^{-1}X^\top y$$

Using part (a), we have $\widehat{\theta}_\lambda = X^\top(XX^\top + \lambda I)^{-1}y$.

Recall that $(XX^\top + \lambda I)$ is positive definite and has real, positive eigenvalues when $\lambda > 0$. The invertibility of this matrix implies a unique solution for $\widehat{\theta}_\lambda$.

(c) Suppose X has the singular value decomposition $U\Sigma V^\top$, where $\Sigma = \text{diag}(s_1, \cdots, s_d)$, $s_i \geq 0$. Show that $\widehat{\theta}_\lambda = VDU^\top y$, where D is a diagonal matrix to be determined.

Solution: Notice that the SVD of $X^\top$ is $V\Sigma U^\top$. By computation, we have

$$(X^\top X + \lambda I) = V\Sigma U^\top U\Sigma V^\top + V(\lambda I)V^\top \tag{1}$$
$$= V(\Sigma^2 + \lambda I)V^\top \tag{2}$$

$(X^\top X + \lambda I)$ can thus be diagonalized into the form $V(\Sigma^2 + \lambda I)V^\top$, with

$$V \,\text{diag}\left(\frac{1}{s_1^2 + \lambda}, \ldots, \frac{1}{s_d^2 + \lambda}\right) V^\top$$

as its inverse. This allows us to write

$$\widehat{\theta}_\lambda = V \,\text{diag}\left(\frac{1}{s_1^2 + \lambda}, \ldots, \frac{1}{s_d^2 + \lambda}\right) V^\top V\Sigma U^\top y = V \,\text{diag}\left(\frac{1}{s_1^2 + \lambda}, \ldots, \frac{1}{s_d^2 + \lambda}\right) \Sigma U^\top y = VDU^\top y$$

where

$$D = \text{diag}\left(\frac{s_1}{s_1^2 + \lambda}, \ldots, \frac{s_d}{s_d^2 + \lambda}\right)$$

(d) Let $\widehat{y}_\lambda = X\widehat{\theta}_\lambda$ be the predictions made by the ridge regressor $\widehat{\theta}_\lambda$. Suppose we have $y = X\theta_* + z$, where $\theta_* \in \mathbb{R}^d$ and $z = \mathcal{N}(0, \sigma^2 I) \in \mathbb{R}^n$ ($\sigma > 0$). Further suppose that X is an orthogonal matrix, that is, $X^\top X = I$.

$\mathbb{E}\|X(\widehat{\theta}_\lambda - \theta_*)\|^2$ is the expected squared difference between the predictions made by the ridge regressor $\widehat{y}_\lambda$ and $X\theta_*$, where the expectation is taken with respect to z ($\|\cdot\|$ denotes ℓ_2 norm).

Show that $\mathbb{E}\|X(\widehat{\theta}_\lambda - \theta_*)\|^2 = \frac{1}{(1+\lambda)^2}\left(\lambda^2\|\theta_*\|^2 + d\sigma^2\right)$.

Solution: First we compute $\widehat{\theta}_\lambda - \theta_*$:

$$\widehat{\theta}_\lambda - \theta_* = (X^\top X + \lambda I)^{-1} X^\top y - \theta_* = ((1+\lambda)I)^{-1} X^\top (X\theta_* + z) - \theta_* = -\frac{\lambda}{1+\lambda}(\theta_*) + \frac{1}{1+\lambda} X^\top z$$

Since X is orthogonal, it is unitary invariant. Therefore,

$$\begin{aligned}
\mathbb{E}\|X(\widehat{\theta}_\lambda - \theta_*)\|^2 &= \mathbb{E}\|\widehat{\theta}_\lambda - \theta_*\|^2 \\
&= \mathbb{E}\Big\| - (\frac{\lambda}{1+\lambda})(\theta_*) + \frac{1}{1+\lambda} X^\top z\Big\|^2 \\
&= \frac{\lambda^2}{(1+\lambda)^2}\|\theta_*\|^2 + \frac{1}{(1+\lambda)^2} d\sigma^2,
\end{aligned}$$

since z is zero mean and $\mathbb{E}\|z\|^2 = d\sigma^2$.

(e) What is the λ^* that you should pick to minimize the prediction error you computed in part (e)? Comment on how d, σ^2, and θ_* affect the optimal choice of the regularization parameter λ.

Solution: Differentiating $\mathbb{E}||X(\widehat{\theta}_\lambda - \theta_*)||^2 = \frac{\lambda^2}{(1+\lambda)^2}||\theta_*||^2 + \frac{1}{(1+\lambda)^2}d\sigma^2$ with respect to λ gives

$$\frac{2(||\theta_*||^2\lambda^* - d\sigma^2)}{(\lambda^* + 1)^3} = 0$$

Therefore, we have that

$$\lambda^* = \frac{d\sigma^2}{||\theta^*||^2}$$

Higher d (more features), higher σ (greater noise), and smaller norm of θ_* (smaller signal) all make us pick larger λ^*.

