2. From Scores to Conditional Probabilities

Problem 2.1. Write $\mathbb{E}_y\left[\ell\left(yf(x)\right) \mid x\right]$ in terms of $\pi(x)$, $\ell(-f(x))$, and $\ell\left(f(x)\right)$.

Solution

$$\mathbb{E}_y\left[\ell\left(yf(x)\right) \mid x\right] = \pi(x)l(f(x)) + (1-\pi(x))l(-f(x))$$

Problem 2.2. Show that the Bayes prediction function $f^*(x)$ for the exponential loss function $\ell\left(y, f(x)\right) = e^{-yf(x)}$ is given by

$$f^*(x) = \frac{1}{2}\ln\left(\frac{\pi(x)}{1-\pi(x)}\right),$$

where we've assumed $\pi(x) \in (0,1)$. Also, show that given the Bayes prediction function f^*, we can recover the conditional probabilities by

$$\pi(x) = \frac{1}{1+e^{-2f^*(x)}}.$$

Solution

$$f^*(x) = argmin \pi(x)e^{-f(x)} + (1-\pi(x))e^{f(x)}$$

Taking the derivative and setting this equal to 0 we get

$$\begin{aligned}
-\pi(x)e^{-f^*(x)} + (1-\pi(x))e^{f^*(x)} &= 0 \\
(1-\pi(x))e^{f^*(x)} &= \pi(x)e^{-f^*(x)} \\
e^{2f^*(x)} &= \frac{\pi(x)}{1-\pi(x)} \\
2f^*(x) &= \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \\
f^*(x) &= \frac{1}{2}\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)
\end{aligned}$$

For the second part observe that

$$\begin{aligned}
e^{2f^*(x)} &= \frac{\pi(x)}{1-\pi(x)} \\
e^{2f^*(x)} - \pi(x)e^{2f^*(x)} &= \pi(x) \\
e^{2f^*(x)} &= \pi(x)(1+e^{2f^*(x)}) \\
\frac{1}{e^{-2f^*(x)}} &= \pi(x)(1+e^{2f^*(x)}) \\
\frac{1}{(1+e^{-2f^*(x)})} &= \pi(x)
\end{aligned}$$

Problem 2.3. Show that the Bayes prediction function $f^*(x)$ for the logistic loss function $\ell\left(y, f(x)\right) = \ln\left(1 + e^{-yf(x)}\right)$ is given by

$$f^*(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

and the conditional probabilities are given by

$$\pi(x) = \frac{1}{1+e^{-f^*(x)}}.$$

Solution

$$f^*(x) = argmin \pi(x)\ln(1+e^{-\hat{y}}) + (1-\pi(x))\ln(1+e^{\hat{y}})$$

Taking the derivative and setting this equal to 0 we get

$$\begin{aligned}
\frac{-\pi(x)e^{-\hat{y}}}{1+e^{-\hat{y}}} + \frac{(1-\pi(x))e^{\hat{y}}}{1+e^{\hat{y}}} &= 0 \\
\frac{-\pi(x)}{1+e^{\hat{y}}} + \frac{(1-\pi(x))e^{\hat{y}}}{1+e^{\hat{y}}} &= 0 \\
-\pi(x) + (1-\pi(x))e^{\hat{y}} &= 0 \\
e^{\hat{y}} &= \frac{\pi(x)}{1-\pi(x)} \\
\hat{y} = f^*(x) &= \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)
\end{aligned}$$

For the second part observe that

$$\begin{aligned}
e^{\hat{y}} &= \frac{\pi(x)}{1-\pi(x)} \\
e^{\hat{y}-\pi(x)e^{\hat{y}}} &= \pi(x) \\
e^{\hat{y}} &= \pi(x)(1+e^{\hat{y})} \\
\frac{e^{\hat{y}}}{1+e^{\hat{y}}} &= \pi(x) \\
\frac{1}{1+e^{-\hat{y}}} = \frac{1}{1+e^{-f^*(x)}} &= \pi(x)
\end{aligned}$$

3. Logistic Regression

Problem 3.1. Show that $n\hat{R}_n(w) = \text{NLL}(w)$ for all $w \in \mathbf{R}^d$. And thus the two approaches are equivalent, in that they produce the same prediction functions.

Solution Recall that

$$\begin{aligned} n\hat{R}_n(w) &= \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i w^T x_i\right)\right) \\ \text{NLL}(w) &= \sum_{i=1}^{n} \left[-y_i' \log \phi(w^T x_i)\right] + (y_i' - 1)\log\left(1 - \phi(w^T x_i)\right) \end{aligned}$$

Let $(x, y) \in D$ and $(x, y') \in D'$
We will consider the cases where $y = 1$ and $y = -1$ separately.

Case 1: $y = -1 \implies y' = 0$
Observe that

$$\log\left(1 + \exp\left(-y_i w^T x_i\right)\right) = \log\left(1 + \exp\left(w^T x_i\right)\right)$$

And that

$$\begin{aligned} \left[-y_i' \log \phi(w^T x_i)\right] + (y_i' - 1)\log\left(1 - \phi(w^T x_i)\right) &= -\log(1 - \phi(w^T x)) \\ &= -\log\left(1 - \frac{1}{1 + e^{-w^T x}}\right) \\ &= -\log(\frac{e^{-w^T x}}{1 + e^{-w^T x}}) \\ &= -\log(\frac{1}{1 + e^{w^T x}}) \\ &= \log(1 + e^{w^T x}) \end{aligned}$$

Thus $\log\left(1 + \exp\left(-y_i w^T x_i\right)\right) = \left[-y_i' \log \phi(w^T x_i)\right] + (y_i' - 1)\log\left(1 - \phi(w^T x_i)\right)$ when $y = -1$

Case 2: $y = 1 \implies y' = 1$
Observe that

$$\log\left(1 + \exp\left(-y_i w^T x_i\right)\right) = \log\left(1 + \exp\left(-w^T x_i\right)\right)$$

And that

$$\begin{aligned} \left[-y_i' \log \phi(w^T x_i)\right] + (y_i' - 1)\log\left(1 - \phi(w^T x_i)\right) &= -\log(\phi(w^T x)) \\ &= \log\left(\frac{1}{1 + e^{-w^T x}}\right) \\ &= \log(1 + e^{-w^T x}) \end{aligned}$$

Thus $\log\left(1+\exp\left(-y_i w^T x_i\right)\right) = \left[-y_i' \log\phi(w^T x_i)\right] + (y_i'-1)\log\left(1-\phi(w^T x_i)\right)$ when $y=1$

We have shown that all terms of the summations are equal. Thus we can conclude that $n\hat{R}_n(w) = \mathrm{NLL}(w)$.

Problem 3.2.1. Show that the expression for LogSumExp is valid.

Solution

$$\begin{aligned}
\log(e^{x_i} + \dots e^{x_n}) &= \log(e^{x^*}(e^{x_1 - x^*} + \dots e^{x_n - x^*})) \\
&= \log(e^{x^*}) + \log(e^{x_1 - x^*} + \dots e^{x_n - x^*}) \\
&= x^* + \log(e^{x_1 - x^*} + \dots e^{x_n - x^*})
\end{aligned}$$

Problem 3.2.2. Show that $\exp(x_i - x^*) \in (0, 1]$ for any i, and thus the exp calculations will not overflow.

Solution Since $x^* = max(x_1, \dots x_n)$, we have $x_i - x^* \leq 0$ for all i. Thus $0 < e^{x_i - x^*} \leq 1$ for all i

Problem 3.2.3. Explain why the log term in our expression $\log\left[e^{x_1-x^*}+\cdots+e^{x_n-x^*}\right]$ will never be "-inf".

Solution Note that there exists at least one x_i such that $x_i = x^*$ and that for such numbers $e^{x_i-x^*} = e^0 = 1$. Thus $e^{x_1-x^*}+\ldots e^{x_n-x^*} > 1$. Thus $\ln(e^{x_1-x^*}+\ldots e^{x_n-x^*}) > e$ and so we don't have to worry about negative infinity.

Problem 3.2.4. Show how to use the numpy function *logaddexp* to correctly compute $\log\left(1+e^{-s}\right)$

Solution np.logaddexp(0,-s)

4. Bayesian Logistic Regression with Gaussian Priors

Problem 4.1. For the dataset $\mathcal{D}'$ described in Section 3.1, give an expression for the posterior density $p(w \mid \mathcal{D}')$ in terms of the negative log-likelihood function

Solution

$$p(w \mid D') \propto p(w)e^{-NLL(w)}$$

Problem 4.2. Suppose we take a prior on w of the form $w \sim \mathcal{N}(0, \Sigma)$. Find a covariance matrix Σ such that MAP estimate for w after observing data $\mathcal{D}'$ is the same as the minimizer of the regularized logistic regression function defined in Section 3.3 (and prove it).

Solution From section 3.3 we have

$$\hat{w} = \text{argmin } \hat{R}_n(w) + \lambda \|w\|^2.$$

Now

$$\begin{aligned} w^* &= \text{argmax } p(w \mid D') \\ &= \text{argmax } p(w) e^{-NLL(w)} \\ &= \text{argmin } -\ln \left(p(w) e^{-NLL(w)} \right) \\ &= \text{argmin } \frac{1}{2} w^T \Sigma^{-1} w + NLL(w) \\ &= \text{argmin } \frac{1}{2n} w^T \Sigma^{-1} w + \hat{R}_n(w) && \text{since } NLL(w) = n\hat{R}_n(w) \end{aligned}$$

Thus

$$\begin{aligned} \lambda w^T w &= \frac{1}{2n} w^T \Sigma^{-1} w \\ \lambda I &= \frac{1}{2n} w^T \Sigma^{-1} \\ \Sigma &= \frac{1}{2n\lambda} I \end{aligned}$$

Problem 4.3. In the Bayesian approach, the prior should reflect your beliefs about the parameters before seeing the data and, in particular, should be independent on the eventual size of your dataset. Following this, you choose a prior distribution $w \sim \mathcal{N}(0, I)$. For a dataset $\mathcal{D}$ of size n, how should you choose λ in our regularized logistic regression objective function so that the minimizer is equal to the mode of the posterior distribution of w (i.e. is equal to the MAP estimator).

Solution

$$\lambda = \frac{1}{2n}$$

6. Coin Flipping Maximum Likelihood

Problem 6.1. Suppose we flip a coin and get the following sequence of heads and tails:

$$\mathcal{D} = (H, H, T)$$

Give an expression for the probability of observing $\mathcal{D}$ given that the probability of heads is θ. That is, give an expression for $p\left(\mathcal{D} \mid \theta\right)$. This is called the **likelihood of θ for the data $\mathcal{D}$**.

Solution

$$p(\mathcal{D} \mid \theta) = \theta^2(1 - \theta)$$

Problem 6.2. How many different sequences of 3 coin tosses have 2 heads and 1 tail? If we toss the coin 3 times, what is the probability of 2 heads and 1 tail? (Answer should be in terms of θ.)

Solution

$$p(2H, 1T) = 3\theta^2(1-\theta)$$

Problem 6.3. More generally, give an expression for the likelihood $p(\mathcal{D} \mid \theta)$ for a particular sequence of flips $\mathcal{D}$ that has n_h heads and n_t tails. Make sure you have expressions that make sense even for $\theta = 0$ and $n_h = 0$, and other boundary cases. You may use the convention that $0^0 = 1$, or you can break your expression into cases if needed.

Solution

$$p(\mathcal{D} \mid \theta) = \theta^{n_h}(1-\theta)^{n_t}$$

Problem 6.4. Show that the maximum likelihood estimate of θ given we observed a sequence with n_h heads and n_t tails is

$$\hat{\theta}_{\text{MLE}} = \frac{n_h}{n_h + n_t}.$$

You may assume that $n_h + n_t \geq 1$. (Hint: Maximizing the log-likelihood is equivalent and is often easier.)

Solution

$$\hat{\theta}_{MLE} = argmax\ \theta^{n_h}(1-\theta)^{n_t} = argmax\ n_h \ln(\theta) + n_t \ln(1-\theta)$$

Taking the derivative and setting it to 0 we get

$$\begin{aligned}
\frac{n_h}{\hat{\theta}} - \frac{n_t}{1-\hat{\theta}} &= 0 \\
n_h(1-\hat{\theta}) - n_t\hat{\theta} &= 0 \\
n_h - n_h\hat{\theta} - n_t\hat{\theta} &= 0 \\
\hat{\theta} &= \frac{n_h}{n_h + n_t}
\end{aligned}$$

7. Coin Flipping: Bayesian Approach with Beta Prior

Problem 7.1. Suppose that our prior distribution on θ is Beta(h, t), for some $h, t > 0$. That is, $p(\theta) \propto \theta^{h-1} (1-\theta)^{t-1}$. Suppose that our sequence of flips $\mathcal{D}$ has n_h heads and n_t tails. Show that the posterior distribution for θ is Beta$(h + n_h, t + n_t)$. That is, show that

$$p(\theta \mid \mathcal{D}) \propto \theta^{h-1+n_h} (1-\theta)^{t-1+n_t}.$$

Solution

$$p(\theta) = \theta^{h-1}(1-\theta)^{t-1}$$

$$\begin{aligned} p(\theta \mid \mathcal{D}) &\propto (\theta)(\mathcal{D} \mid \theta) \\ &= \theta^{h-1}(1-\theta)^{t-1}\theta^{n_h}(1-\theta)^{n_t} \\ &= \theta^{h+n_h-1}(1-\theta)^{t+n_t-1} \end{aligned}$$

Problem 7.2. Give expressions for the MLE, the MAP, and the posterior mean estimates of θ.

Solution

$$
\begin{aligned}
MLE &= \frac{n_h}{n_h + n_t} \\
MAP &= \frac{h + n_h - 1}{h + t + n_h + n_t - 2} \\
PosteriorMean &= \frac{h_{nh}}{h + t + n_h + n_t}
\end{aligned}
$$

Problem 7.3. What happens to $\hat{\theta}_{\text{MLE}}$, $\hat{\theta}_{\text{MAP}}$, and $\hat{\theta}_{\text{POSTERIOR MEAN}}$ as the number of coin flips $n = n_h + n_t$ approaches infinity?

Solution They all approach the same value. $MLE = MAP = PosteriorMean = \frac{n_h}{n_h+n_t}$

Problem 7.4. The MAP and posterior mean estimators of θ were derived from a Bayesian perspective. Let's now evaluate them from a frequentist perspective. Suppose θ is fixed and unknown. Which of the MLE, MAP, and posterior mean estimators give **unbiased** estimates of θ, if any?

Solution The MLE is always unbiased and the MAP is unbiased only if $h = t = 1$.

$$E(MLE) = \frac{1}{n}E(n_h) = \frac{1}{n}n\theta = \theta$$

$$E(MAP) = \frac{h-1}{n+h+t-2} + \frac{n\theta}{n+h+t-2} = \theta \qquad \text{when } h = t = 1$$

$$E(PosteriorMean) = \frac{h}{n+h+t} + \frac{n\theta}{n+h+t} \neq \theta \qquad \text{since } h, t > 0$$

Problem 7.5. Suppose somebody gives you a coin and asks you to give an estimate of the probability of heads, but you can only toss the coin 3 times. You have no particular reason to believe this is an unfair coin. Would you prefer the MLE or the posterior mean as a point estimate of θ? If the posterior mean, what would you use for your prior?

Solution I would prefer the posterior mean. My belief is that most coins are close to fair. With a sample size of only 3, the closest to $\theta = 0.5$ we could get is $\theta = \frac{1}{3}$ or $\theta = \frac{2}{3}$. I think that it is very unlikely that even a trick coin could be that far from fair. For a prior, I would choose a Beta distribution centered at 0.5 with low variance, something like $Beta(100, 100)$.

8. Hierarchical Bayes for Click-Trhough Rate Estimation

Problem 8.1.1. Give an expression for $p(\mathcal{D}_i \mid \theta_i)$, the likelihood of $\mathcal{D}_i$ given the probability of click θ_i, in terms of θ_i, x_i and n_i.

Solution

$$p(D_i \mid \theta_i) = \theta^{x_i}(1-\theta)^{n_I - x_i}$$

Problem 8.1.2. We will take our prior distribution on θ_i to be $\text{Beta}(a, b)$. The corresponding probability density function is given by

$$p(\theta_i) = \text{Beta}(\theta_i; a, b) = \frac{1}{B(a,b)} \theta_i^{a-1} \left(1 - \theta_i\right)^{b-1},$$

where $B(a, b)$ is called the Beta function. Explain (without calculation) why we must have

$$\int \theta_i^{a-1} \left(1 - \theta_i\right)^{b-1} \, d\theta_i = B(a, b).$$

Solution The area under the PDF must equal 1.

Problem 8.1.3. Give an expression for the posterior distribution $p(\theta_i \mid \mathcal{D}_i)$. In this case, include the constant of proportionality. In other words, do not use the "is proportional to" sign $\propto$ in your final expression.

Solution Note that

$$p(\theta_i \mid D_i) \propto \theta_i^{a+x_i-1}(1-\theta_i)^{n_i+b-x_i-1}$$

Now

$$\int \theta_i^{a+x_i-1}(1-\theta_i)^{n_i+b-x_i-1} d\theta_i = B(a+x_i, n_i+b-x_i)$$

Thus, in order to have a valid PDF,

$$p(\theta_i \mid D_i) = \frac{1}{B(a+x_i, n_i+b-x_i)} \theta_i^{a+x_i-1}(1-\theta_i)^{n_i+b-x_i-1}$$

Problem 8.1.4. Give a closed form expression for $p(\mathcal{D}_i)$, the marginal likelihood of $\mathcal{D}_i$, in terms of the a, b, x_i, and n_i. You may use the normalization function $B(\cdot, \cdot)$ for convenience, but you should not have any integrals in your solution.

Solution

$$\begin{aligned} p(D_i) &= \int p(D_i \mid \theta_i) p(\theta_i) d\theta_i \\ &= \int \frac{1}{B(a,b)} \theta_i^{a+x_i-1} (1-\theta_i)^{n_i+b-x_i-1} d\theta_i \\ &= \frac{B(a+x_i, n_i+b-x_i)}{B(a,b)} \end{aligned}$$

Problem 8.1.4. The maximum likelihood estimate for θ_i is x_i/n_i. Let $p_{\mathrm{MLE}}(\mathcal{D}_i)$ be the marginal likelihood of $\mathcal{D}_i$ when we use a prior on θ_i that puts all of its probability mass at x_i/n_i. Note that

$$\begin{aligned} p_{\mathrm{MLE}}(\mathcal{D}_i) &= p\left(\mathcal{D}_i \mid \theta_i = \frac{x_i}{n_i}\right) p\left(\theta_i = \frac{x_i}{n_i}\right) \\ &= p\left(\mathcal{D}_i \mid \theta_i = \frac{x_i}{n_i}\right). \end{aligned}$$

Explain why, or prove, that $p_{\mathrm{MLE}}(\mathcal{D}_i)$ is larger than $p(\mathcal{D}_i)$ for any other prior we might put on θ_i.

Solution By definition, $\theta_{MLE} = \frac{x_i}{n} = \text{argmax}\, p(D_i \mid \theta)$.
Thus $p(D_i \mid \theta_{MLE}) \geq p(D_i \mid \theta) \;\; \forall \;\; \theta$
Therefore $p(\mathcal{D}_i) = \int p\left(\mathcal{D}_i \mid \theta_i\right) p(\theta_i)\, d\theta_i$ is maximized when

$$p(\theta) = \begin{cases} 1 & \theta = \frac{x_i}{n} \\ 0 & \text{otherwise} \end{cases}$$

Problem 8.1.5. Explain what's happening to the prior as we continue to increase the likelihood.

Solution The prior's mode approached the MLE of θ and the variance approaches 0.