Homework 5: Conditional Probability Models

August 28, 2021

1 Introduction

No question here.

2 From Scores to Conditional Probabilities

1. We have: $y \in \{-1, 1\}$ and $\pi(x) = \mathbb{P}(y = 1|x)$.
   The conditional expectation of $l(yf(x))$ is:

$$\begin{aligned} \mathbb{E}_y[l(yf(x))|x] &= l(f(x))\mathbb{P}(y = 1|x) + l(-f(x))\mathbb{P}(y = -1|x) \\ &= l(f(x))\pi(x) + l(-f(x))[1 - \pi(x)] \end{aligned}$$

2. If $l(y(f(x)) = e^{-yf(x)}$, we have:

$$\begin{aligned} \mathbb{E}_y[l(yf(x))] &= e^{-yf(x)}\pi(x) + e^{yf(x)}[1 - \pi(x)] \\ &\geq 2\sqrt{\pi(x)[1 - \pi(x)} \end{aligned}$$

   The equality holds only when $e^{-yf(x)}\pi(x) = e^{yf(x)}[1 - \pi(x)]$, so $f^*(x) = \frac{1}{2} \ln \frac{\pi(x)}{[1 - \pi(x)]}$.
   If $f^*(x)$ is given, by simple algebraic operation, $\pi(x) = \dfrac{1}{1 + e^{-2f^*(x)}}$.

3. If $l(y(f(x)) = \ln(1 + e^{-yf(x)})$, the conditional expectation is:

$$\begin{aligned} \mathbb{E}_y[l(yf(x))|x] &= \ln(1 + e^{-f(x)})\pi(x) + \ln(1 + e^{f(x)})[1 - \pi(x)] \\ \frac{\partial}{\partial f}\mathrm{E} &= \frac{-e^{-f(x)}\pi(x)}{1 + e^{-f(x)}} + \frac{e^{f(x)}[1 - \pi(x)]}{1 + e^{f(x)}} \\ &= \frac{e^{f(x)}[1 - \pi(x)] - \pi(x)}{1 - e^{f(x)}} \end{aligned}$$

$f^*(x)$ satisfies $\dfrac{\partial}{\partial f^*}E = 0$, so:

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

4. If $l(y, f(x)) = \max\{0, 1 - yf(x)\}$, the conditional expectation is:

   $$\mathbb{E}_y[l(y, f(x))|x] = \max\{0, 1 - f(x)\}\pi(x) + \max\{0, 1 + f(x)\}[1 - \pi(x)]$$

   Denote $\mathbb{E}_y[l(y, f(x))|x]$ as E, then:

   $$\begin{aligned} \mathrm{E} &= \begin{cases} [1+f(x)][1-\pi(x)] & f(x) > 1 \\ 1 + f(x) - 2f(x)\pi(x) & -1 \le f(x) \le 1 \\ [1-f(x)]\pi(x) & f(x) < -1 \end{cases} \\ \frac{\partial}{\partial f}\mathrm{E} &= \begin{cases} 1-\pi(x) & f(x) > 1 \\ 1-2\pi(x) & -1 < f(x) < 1 \\ -\pi(x) & f(x) < -1 \end{cases} \end{aligned}$$

   $\pi(x)$ represents a probability, so $1-\pi(x) \ge 0, -\pi(x) \le 0$. Hence, E reaches its maxima at -1 when $1 - 2\pi(x) < 0$, at 1 when $1 - 2\pi(x) > 0$ and at any
   points in $[-1, 1]$ when $\pi(x) = \dfrac{1}{2}$. Therefore, we have:

   $$f^*(x) = \mathrm{sign}[1 - 2\pi(x)]$$

3 Logistic Regression

3.1 Equivalence of ERM and probabilistic model

1. The ERM of the logistic loss function is:

   $$\hat{R}(w) = \frac{1}{n}\sum_{i=1}^{n} \ln(1 + e^{-y_i w^T x_i})$$

   The negative log likelihood of logistic probability is:

$$
\begin{aligned}
NLL(w) &= -\sum_{i=0}^{n} \ln \frac{1}{1+e^{-y_i w^T x_i}} \\
&= \sum_{i=0}^{n} \ln(1+e^{-y_i w^T x_i})
\end{aligned}
$$

Hence, minimizing the ERM is equivalent with minimizing negative log likelihood.

3.2 Numeric overflow and log-sum-exp trick

1. Let $x^* = \max_i x_i, i = 1, ..., n$, then:

$$
\begin{aligned}
\ln(\sum_{i=1}^{n} e^{x_i}) &= \ln(e^{x^*} \sum_{i=1}^{n} e^{x_i - x^*}) \\
&= x^* + \ln(\sum_{i=1}^{n} e^{x_i - x^*})
\end{aligned}
$$

2. Because $x^* = \max_i x_i, i = 1, ..., n$, $x_i - x^{^*} \leq 0$ and $e^{x_i - x^*} \leq 1$.
   Hence, the sum of exponential calculation will not overflow.

3. At least one of $x_i - x^*$ is 0 according to the definition of x^*, so $1 < \sum_{i=1}^{n} e^{x_i - x^*} \leq n$.
   Therefore, the logrithm calculation will not overflow.

4. Just use the numpy logaddexp() function in this way: numpy.logaddexp(0, -s).

3.3 Regularized Logistic Regression

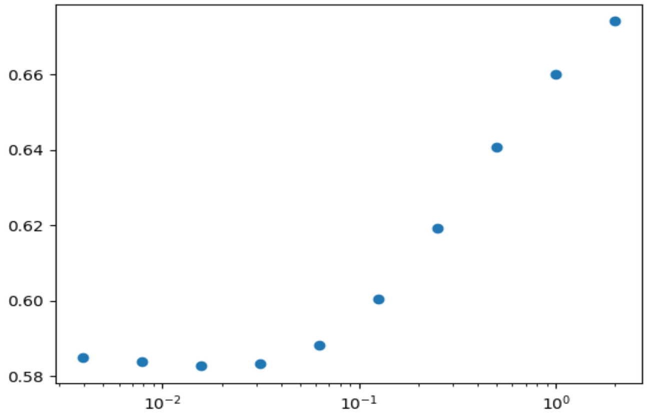
1. Denote $\ln(1 + e^{-yw^T x})$ as $f(w)$, so:

$$
\begin{aligned}
\nabla_w f(w) &= \frac{-yx}{1+e^{yw^T x}} \\
\mathbf{H}_w f(w) &= \frac{y^2 e^{yw^T x}}{(1+e^{yw^T x})^2} x x^T
\end{aligned}
$$

   The 2nd-order derivative w.r.t. w is positive semi-definite, so $f(w)$ is convex function.

2. Programming question.

3. Programming question.
4. The values of negative log-likelihood against regularization parameters are shown below:



4 Bayesian Logistic Regression with Gaussian Prior

1. Given the dataset $\mathcal{D}^{'}$, its negative log-likelihood $NLL_{\mathcal{D}'}(w)$ and prior density $p(w)$, the posterior density is:

$$p(w|\mathcal{D}^{'}) \propto \exp[-NLL_{\mathcal{D}'}(w)] \cdot p(w)$$

2. If $p(w) \sim \mathcal{N}(0, \Sigma)$ and w^* is the MAP estimate of w:

$$\begin{aligned}
w^* &= \arg\max_w p(w|\mathcal{D}^{'}) \\
&= \arg\max_w \exp[-NLL_{\mathcal{D}'}(w)] \cdot p(w) \\
&= \arg\max_w -NLL_{\mathcal{D}'}(w) + \ln p(w) \\
&= \arg\max_w -NLL_{\mathcal{D}'}(w) - w^T\Sigma^{-1}w/2 \\
&= \arg\min_w NLL_{\mathcal{D}'}(w) + w^T\Sigma^{-1}w/2
\end{aligned}$$

If $\Sigma = \dfrac{1}{2n\lambda}I, \Sigma^{-1} = 2n\lambda I$, so:

$$w^* = \arg\min_w \frac{1}{n}NLL_{\mathcal{D}'}(w) + \lambda w^T w$$

which is exactly as same as the form of regularied logistic regression.

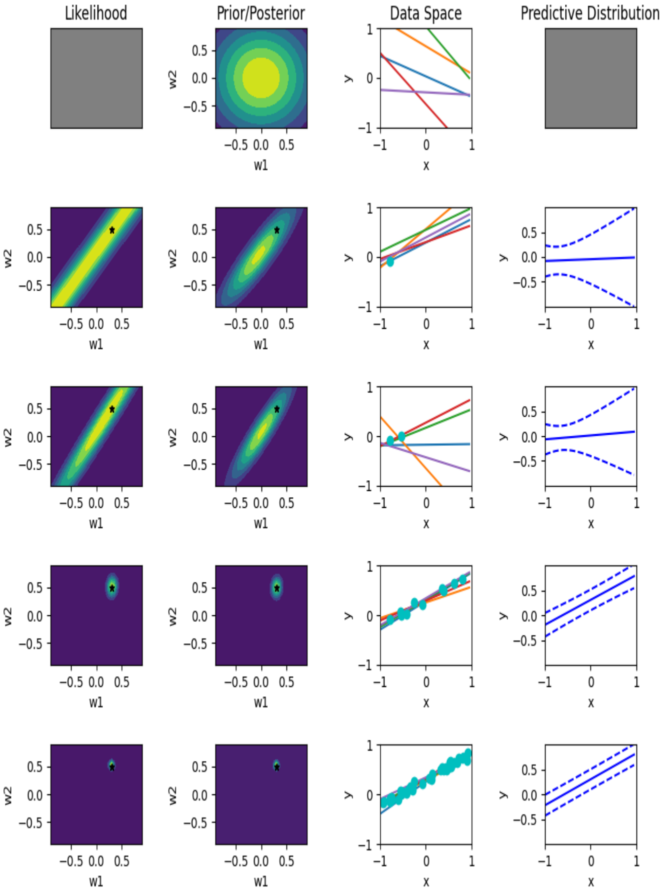
3. As the result got in 4.2, $\Sigma = \dfrac{1}{2n\lambda} I$.

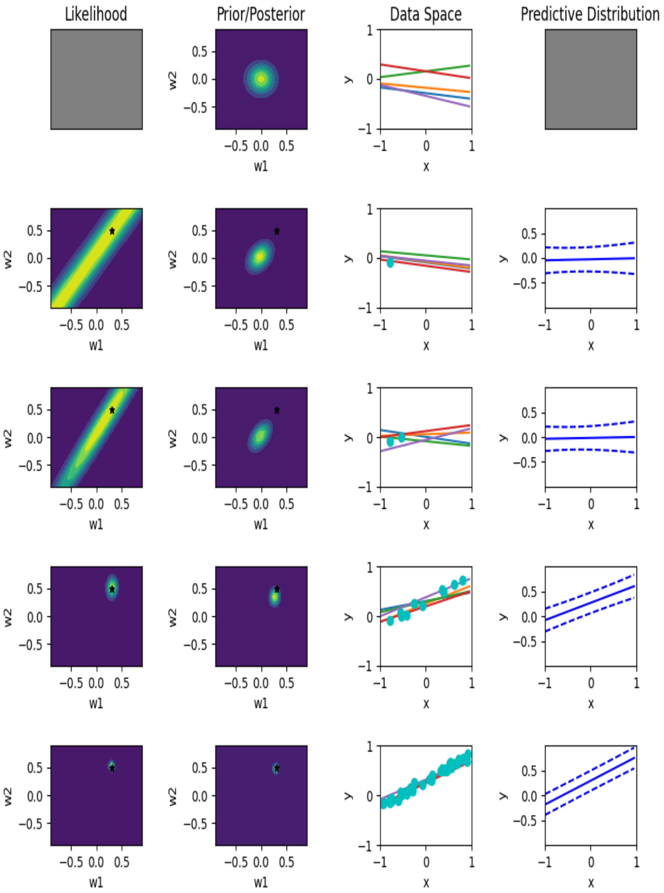
   If $w \sim \mathcal{N}(0, I)$, $2n\lambda = 1$,so $\lambda = \dfrac{1}{2n}$.

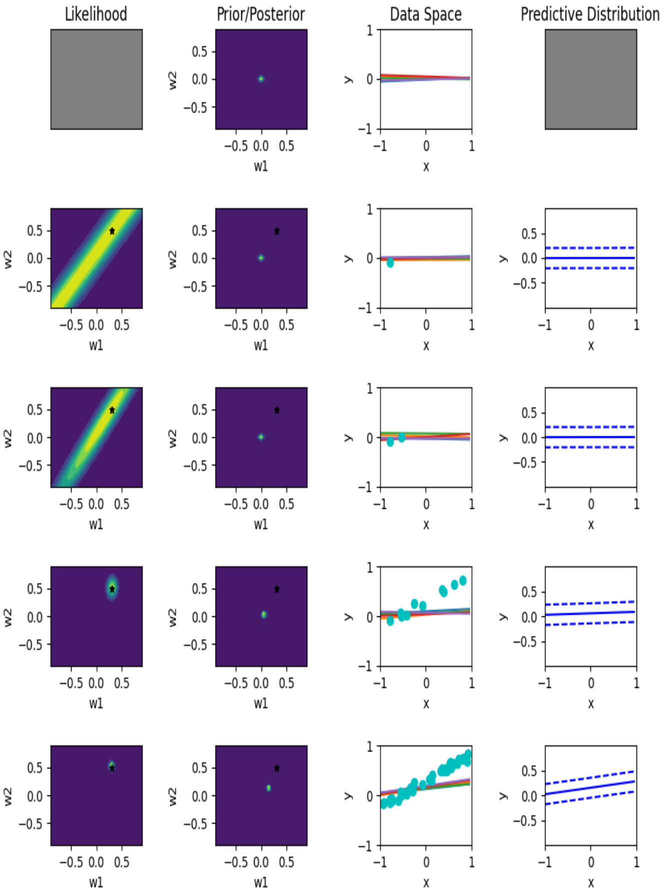
   In this way, regularized logistic regression is equivalent with MAP estimate of bayesian regression.

5 Bayesian Linear Regression

1. Programming question.
2. Programming question.
3. Programming question.
4. The results of bayesian regression with three different settings are shown below:


Likelihood
Prior/Posterior
Data Space
Predictive Distribution
w1
w2
x
y
−0.5
0.0
0.5
−1
0
1


Likelihood
Prior/Posterior
Data Space
Predictive Distribution
w1
w2
x
y
−0.5
0.0
0.5
−1
0
1



5. Suppose y_i is sampled from $\mathcal{N}(w^T x_i, 2\sigma^2)$, denote w^* as the MAP estimate of w.
If the prior is $\mathcal{N}(0, \frac{1}{2}I)$, then:

$$
\begin{aligned}
w^* &= \arg\max_w \mathsf{p}(\mathcal{D}|w)\mathsf{p}(w) \\
&= \arg\max_w \ln[\mathsf{p}(\mathcal{D}|w)\mathsf{p}(w)] \\
&= \arg\min_w \sum_{i=1}^{n} \frac{(y_i - w^T x_i)^2}{2\sigma^2} + \frac{w^t(\frac{1}{2}I)^{-1}w}{2} \\
&= \arg\min_w \sum_{i=1}^{n} (y_i - w^T x_i)^2 + 2\sigma^2 w^T w
\end{aligned}
$$

which is equivalent with regularized ridge regression. So just set regularization coefficient as $2\sigma^2$.

6 [Optional] Coin Flipping: Maximum Likelihood

1. If the probability of head is θ, then:

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

2. The probability of 2 heads and 1 tail is:

$$C_3^2\theta^2(1-\theta) = 3\theta^2(1-\theta)$$

3. The probability of η_h heads and η_t tails is:

$$C_{\eta_h+\eta_t}^{\eta_t}\theta^{\eta_h}(1-\theta)^{\eta_t}$$

4. To maximize the probability of the above question, denote $p(\theta) = \theta^{\eta_h}(1-\theta)^{\eta_t}$:

$$
\begin{aligned}
\ln[p(\theta)] &= \eta_h \ln\theta + \eta_t \ln(1-\theta) \\
\frac{d}{d\theta}\ln[p(\theta)] &= \frac{\eta_h}{\theta} - \frac{\eta_t}{1-\theta} \\
&= \frac{\eta_h - (\eta_t + \eta_h)\theta}{\theta(1-\theta)} \\
\theta^* &= \frac{\eta_h}{\eta_t + \eta_h}
\end{aligned}
$$

7 [Optional] Coin Flipping: Bayesian Approach with Beta Prior

1. If $\theta \sim Beta(h, t)$ and the squence of flips $\mathcal{D}$ has n_h heads and n_t tails, then:

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

   So $(\theta|\mathcal{D}) \sim Beta(n_h + h, n_t + t)$.

2. For MAP estimate of θ:

$$\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\theta} \theta^{n_h+h-1}(1-\theta)^{n_t+t-1} \\
&= \frac{n_h + h - 1}{n_h + n_t + t + h - 2}
\end{aligned}$$

   For MLE estimate of θ:

$$\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_{\theta} \theta^{n_h}(1-\theta)^{n_t} \\
&= \frac{n_h}{n_h + n_t}
\end{aligned}$$

   For posterior mean estimate of θ:

$$\begin{aligned}
\hat{\theta}_{PM} &= \mathbb{E}(\theta|\mathcal{D}) \\
&= \frac{B(n_t + t + 1, n_h + h)}{B(n_t + t, n_h + h)} \\
&= \frac{\Gamma(n_t + t + 1)\Gamma(n_h + h)}{\Gamma(n_t + t + n_h + h + 1)} \Bigg/ \frac{\Gamma(n_t + t)\Gamma(n_h + h)}{\Gamma(n_t + t + n_h + h)} \\
&= \frac{n_t + t}{n_t + t + n_h + h}
\end{aligned}$$

3. When the number of coin flippings goes infinity, all the estimates of θ above is the true probability of heads.

4. $\hat{\theta}_{MLE}$ is the unbiased estimate of θ.

$$
\begin{aligned}
n_h &= \sum_{i=1}^{n} 1(X_i = 1) \\
n_t + n_h &= n \\
\mathbb{E}(X_i) &= \theta
\end{aligned}
$$

$X_i, i = 1, ..., n$ are independent variables so the expectation of $\hat{\theta}_{MLE}$ is:

$$\mathbb{E}(\hat{\theta}_{MLE}) = \frac{\theta n}{n} = \theta$$

5. I prefer PM estimate because it is more flexible to ultilize prior knowledge. For this question, because there is no particular reason to believe the coin is unfair, the prior must be symmetric along $\theta = \frac{1}{2}$.
   I may choose $Beta(0, 0)$ as the prior.