Optimization and Machine Learning, Spring 2020
Homework 2
(Due Wednesday, Apr. 1 at 11:59pm (CST))

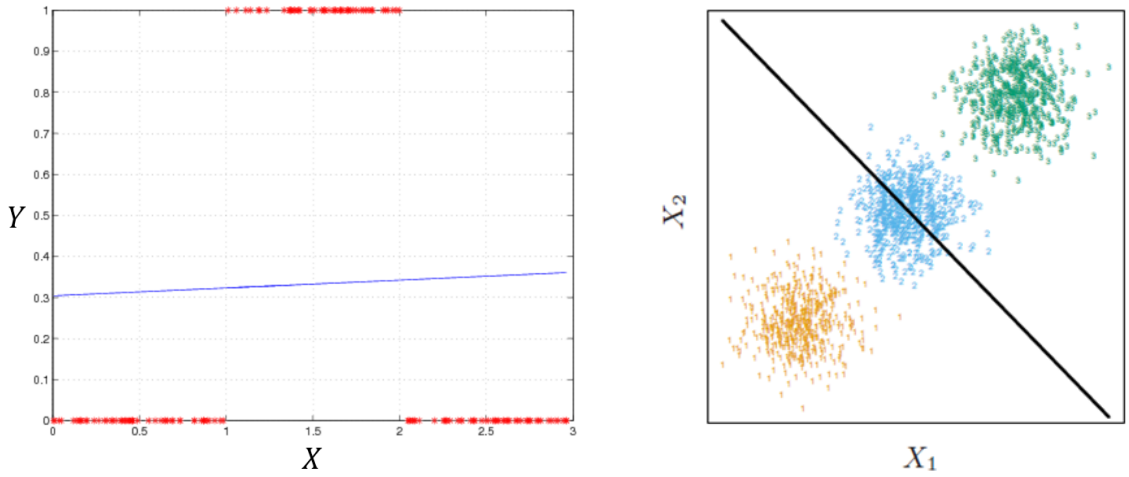
April 5, 2020

1. Suppose that we have N training samples, in which each sample is composed of p input variable and one categorical response with K states.

   (a) Please define this multi-class classification problem, and solve it by ridge regression. (4 points)
   Input:
   $$\mathbf{X} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix},$$
   where x_i is the i-th observation with p parameter.
   Output:
   $$\mathbf{Y} = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix},$$
   where $y_i = (0, \cdots, 0, 1, 0, \cdots, 0)^T$, with its k-th element being 1, indicating that the k-th class is associated with the i-th observation x_i.
   By minimizing the following objective function,
   $$||\mathbf{XB} - \mathbf{Y}||_F^2 + \lambda||\mathbf{B}||_F^2,$$
   we can get the solution $\mathbf{B} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{X}^T\mathbf{Y}$, where $\lambda > 0$ denotes the regularization parameter.

   (b) Please make the prediction of a testing sample $x \in \mathbb{R}^p$ based on your model in (a). (3 points)
   The prediction is made by
   $$\hat{y} = \arg\max_k \hat{f}_k(x),$$
   where $\hat{f}_k(x)$ is the k-th element of
   $$\hat{f}(x) = x^T\mathbf{B} = \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \vdots \\ \hat{f}_K(x) \end{pmatrix}.$$

   (c) Is there any limitation on your model? If yes, please explain the problem by drawing a picture. (3 points)
   The masking problem:



(a) Two classes (b) Three classes

Figure 1: Illustion of the masking problem in linear regression for classification.

(d) Can you propose a model to overcome this limitation? If yes, please derive the decision boundary between an arbitrary class-pair. (5 points)
Linear discriminant analysis (LDA). In LDA, the decision boundary between two arbitrary classes A and B is

$$\mathbf{x}^T\hat{\Sigma}^{-1}(\hat{\mu}_A - \hat{\mu}_B) + \left(\ln(\frac{\Pr(A)}{\Pr(B)}) - \frac{\hat{\mu}_A\hat{\Sigma}^{-1}\hat{\mu}_A - \hat{\mu}_B\hat{\Sigma}^{-1}\hat{\mu}_B}{2}\right) = 0.$$

(e) Can you revise your model in (d) by strength or weaken its assumptions? If yes, please tell the difference between your models in (d) and (e). (5 points)
We can use quadratic discriminative analysis (QDA) for classification.
Difference:

- LDA and QDA both assume that the class conditional probability distributions are normally distributed with different means μ_k, but LDA is different from QDA in that it requires all of the distributions to share the same covariance matrix Σ and QDA requires all of the distribution to have different covariance matrix Σ_k.
- The decision boundary is linear in LDA and quadratic in QDA.
- The number of estimated parameters is $p \times (K + p)$ in LDA and $K \times p \times (p + 1)$ in QDA.

2. Given an random variable, we have N i.i.d. observations by repeated experiments.

(a) If the variable is boolean, please calculate the log-likelihood function. (4 points)
Let X be a boolean random variable which can take either value 1 or 0, and let $\theta = \Pr(X = 1)$ refer to the true. Given an i.i.d dataset $\mathcal{D} = \{x_1, x_2, ..., x_N\}$, we observe $X = 1$ in a total of α_1 times, and $X = 0$ in a total of α_0 times. We denote the likelihood function by $L(\theta) = \Pr(\mathcal{D}|\theta)$:

$$L(\theta) = \Pr(\mathcal{D}|\theta) = \prod_{i=1}^{N} \Pr(X = x_i|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}, \quad x_i \in \{0, 1\}.$$

By taking log on both sides, we have the log-likelihood function, i.e.,

$$\ell(\theta) = \ln L(\theta) = \alpha_1 \ln\theta + \alpha_0 \ln(1-\theta).$$

(b) If the variable is categorical, please calculate the log-likelihood function. (4 points)
Suppose that $X \in \{1, 2, ..., K\}$, and $\theta = \{\theta_1, ..., \theta_K\}$, in which the k-th element $\theta_k = \Pr(X = k)$. The likelihood function $L(\theta) = \Pr(\mathcal{D}|\theta)$ is then derived in the similar way with (a),

$$L(\theta) = \Pr(\mathcal{D}|\theta) = \prod_{i=1}^{N} \Pr(X = x_i|\theta) = \prod_{k=1}^{K} \theta_k^{\alpha_k}, \quad x_i \in \{1, ..., K\}.$$

where α_k counts the number of $x_i = k$ in $\mathcal{D}$, $\forall i, k$. Thus, the log-likelihood function is calculated by

$$\ell(\theta) = \ln L(\theta) = \sum_{k=1}^{K} \alpha_k \ln \theta_k.$$

(c) If the variable is continuous and follows Gaussian distribution, please calculate the log-likelihood function. (5 points)
Let X be a Gaussian random variable parameterized by mean μ and variance σ. Its PDF is given by

$$\mathcal{N}\left(X|\mu, \sigma^2\right) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Under the i.i.d. assumption, the likelihood function $L(\mu, \sigma)$ is expressed as follows,

$$L(\mu, \sigma) = \Pr(\mathcal{D}|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2},$$

and the log-likelihood function becomes

$$\ell(\mu, \sigma) = \ln L(\mu, \sigma) = \frac{N}{2} \ln \frac{1}{2\pi\sigma^2} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

(d) Please discuss the difference between Maximum Likelihood Estimation (MLE) and Maximum a Posterior (MAP) estimation based on ONE of your results in (a), (b) and (c). (7 points)
MLE seeks an estimation of θ that maximizes the conditional probability $\Pr(\mathcal{D}|\theta)$; in contrast, MAP aims to estimate θ by maximizing its posterior $\Pr(\theta|\mathcal{D})$, leading to $\Pr(\theta|\mathcal{D}) \propto \Pr(\mathcal{D}|\theta)\Pr(\theta)$.
We can also see the difference by analyzing the results of (a), in which MLE produces

$$\hat{\theta}^{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0},$$

while MAP gives rise to

$$\hat{\theta}^{MAP} = \frac{\alpha_1 + \beta_1 - 1}{(\alpha_1 + \beta_1 - 1) + (\alpha_0 + \beta_0 - 1)},$$

with the Beta prior $Beta(\beta_1, \beta_0)$.

3. Given the input variables $X \in \mathbb{R}^p$ and a response variable $Y \in \{0, 1\}$, the Expected Prediction Error (EPE) is defined by

$$\text{EPE} = \mathbb{E}[L(Y, \hat{Y}(X))],$$

where $\mathbb{E}(\cdot)$ denotes the expectation over the joint distribution $\Pr(X, Y)$, and $L(Y, \hat{Y}(X))$ is a loss function measuring the difference between the estimated $\hat{Y}(X)$ and observed Y.

(a) Given the zero-one loss

$$L(k, \ell) = \begin{cases} 1 & \text{if } k \neq \ell \\ 0 & \text{if } k = \ell, \end{cases}$$

please derive the Bayes classifier $\hat{Y}(x) = \text{argmax}_{k \in \{0,1\}} \Pr(Y = k|X = x)$ by minimizing EPE. (2 points)
Without loss of generality, we consider $Y \in \{1, 2, ..., M\}$, and rewrite EPE as follows

$$\begin{aligned} \text{EPE} &= \mathbb{E}[L(Y, \hat{Y}(X)] \\ &= \int_x \left[\sum_{m=1}^{M} L\left(Y = m, \hat{Y}(x)\right) \Pr(Y = m|X = x) \right] dx \\ &= \int_x \left[1 - \Pr\left(Y = \hat{Y}(x)|X = x\right) \right] dx. \end{aligned}$$

Therefore,

$$\hat{Y}(x) = \text{argmin EPE} = \text{argmax}_{m \in \{1,...,M\}} \Pr(Y = m|X = x).$$

(b) Please define a function which enables to map the range of an arbitrary linear function to the range of a probability. (2 points)
Given an arbitrary liner function,

$$f(X) = \beta_0 + X^\top \beta \in (-\infty, +\infty),$$

the required function can be defined by

$$\Pr(Y|X) = \frac{\exp(f(X))}{1 + \exp(f(X))} \in (0, 1).$$

(c) Based on the function you defined in (b), please approximate the Bayes classifier in (a) by a linear function between X and Y, and derive its decision boundary. (4 points)
Based on (a), we have

$$\Pr(Y = 0|X) = \frac{\exp(f(X))}{1 + \exp(f(X))},$$
$$\Pr(Y = 1|X) = 1 - \Pr(Y = 0|X) = \frac{1}{1 + \exp(f(X))}.$$

Thus, using the Bayes classifier in (a), we assign the label $Y = 0$ if the following conditions hold:

$$\begin{aligned} & 1 < \frac{P(Y = 0|\boldsymbol{X})}{P(Y = 1|\boldsymbol{X})} \\ \Longrightarrow\ & 0 < \ln \exp(f(\boldsymbol{x})) \\ \Longrightarrow\ & 0 < f(\boldsymbol{x}), \end{aligned}$$

and assign $Y = 1$ otherwise. Hence, we obtain the linear decision boundary $\{X|\beta_0 + X^\top \beta = 0\}$.

(d) If each element of X is boolean, please show how many independent parameters are needed in order to estimate $\Pr(Y|X)$ directly; and is there any way to reduce its number? If yes, please describe your way mathematically. (4 points)
Given $X \in \{0, 1\}^p$ and $Y \in \{0, 1\}$, we need 2^p parameters to estimate $\Pr(Y = 1|X)$, and another 2^p parameters to estimate $\Pr(Y = 0|X)$. However, because of $\Pr(Y = 0|X) = 1 - \Pr(Y = 1|X)$, there are 2^p independent parameters in total.
To reduce the number of parameters, conditional independent assumption is applied, such that

$$\begin{aligned} \Pr(Y|X) &\propto \Pr(X_1, \ldots, X_p|Y)\Pr(Y) \\ &= \prod_{j=1}^{p} \Pr(X_j|Y)\Pr(Y), \end{aligned}$$

according to which we only need to estimate $2p$ independent parameters.

(e) Based on your results in (d) and the Bayes theorem, please develop a classifier with a linear number of parameters w.r.t. p, and estimate these parameters by MLE. (5 points)
Naive Bayes:
Based on the results in (d) and the Bayes theorem, we immediately obtain the naive Bayes classifier:

$$\hat{y} = \text{argmax}_{m \in \{0,1\}} \Pr(Y = m) \prod_{j=1}^{p} \Pr\left(X_j = k|Y = m\right), \quad k \in \{0, 1\}$$

MLE:
According to our discussion in Q2 (a) and (d), $\Pr(X_j = k|Y = m)$ $(k, m \in \{0, 1\})$ is estimated on a training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ by

$$\widehat{\Pr}\left(X_j = k|Y = m\right) = \frac{\sum_{i=1}^N \mathbf{1}_{x_{ij}=k} \mathbf{1}_{y_i=m}}{\sum_{i=1}^N \mathbf{1}_{y_i=m}},$$
$$\widehat{\Pr}\left(Y = m\right) = \frac{\sum_{i=1}^N \mathbf{1}_{y_i=m}}{N},$$

where $\mathbf{1}_{(\cdot)}$ denotes the indicator function.

(f) Please find at least three different points between your developed models in (c) and (e). (3 points)
Difference:

- Naive Bayes in (e) assumes that the random variables are conditional independent given Y, whereas logistic regression in (c) does not hold such assumption.
- Naive Bayes in (e) estimates the parameters of $\Pr(X|Y)$ and $\Pr(Y)$, whereas logistic regression in (c) choose to directly approximate $\Pr(Y|X)$ by a linear function.
- Naive Bayes is a generative model since it models $\Pr(X, Y)$, while logistic regression is a discriminative model as it approximates $\Pr(Y|X)$.
- Two models will converge toward their asymptotic accuracies at different rates.

4. Consider 12 labeled data points sampled from three distinct classes:

$$\text{Class 0}: \begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}-2\\0\end{bmatrix}, \begin{bmatrix}5\\3\end{bmatrix}, \begin{bmatrix}-3\\-5\end{bmatrix} \qquad \text{Class 1}: \begin{bmatrix}\sqrt{2}\\\sqrt{2}\end{bmatrix}, \begin{bmatrix}-\sqrt{2}\\\sqrt{2}\end{bmatrix}, \begin{bmatrix}4\sqrt{2}\\-\sqrt{2}\end{bmatrix}, \begin{bmatrix}-4\sqrt{2}\\-\sqrt{2}\end{bmatrix}, \qquad \text{Class 2}: \begin{bmatrix}3\\5\end{bmatrix}, \begin{bmatrix}-1\\3\end{bmatrix}, \begin{bmatrix}8\\6\end{bmatrix}, \begin{bmatrix}0\\-2\end{bmatrix}$$

(a) For each class $C \in [0, 1, 2]$, compute the class sample mean μ_C, the class sample covariance matrix Σ_C, and the estimate of the prior probability π_C that a point belongs to class C. (6 points)
Solution:

$$\text{Class 0}: \mu_0 = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma_0 = \begin{bmatrix}\frac{38}{3} & 10\\10 & \frac{38}{3}\end{bmatrix}, \pi_0 = \frac{1}{3},$$
$$\text{Class 1}: \mu_1 = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma_1 = \begin{bmatrix}\frac{68}{3} & 0\\0 & \frac{8}{3}\end{bmatrix}, \pi_1 = \frac{1}{3},$$
$$\text{Class 2}: \mu_2 = \begin{bmatrix}2.5\\3\end{bmatrix}, \Sigma_2 = \begin{bmatrix}\frac{49}{3} & 10\\10 & \frac{38}{3}\end{bmatrix}, \pi_1 = \frac{1}{3}.$$

(b) Suppose that we apply LDA to classify the data given in part (a). Will this get the good decision boundary? Briefly explain your answer. (4 points)
The discriminant functions for classes 0 and 1 would have the exact same mean, so there would be no decision boundary between them.

5. We have two classes, named N for normal and E for exponential. For the former class ($Y = N$), the prior probability is $\pi_N = P(Y = N) = \frac{\sqrt{2\pi}}{1+\sqrt{2\pi}}$ and the class conditional $P(X|Y = N)$ has the normal distribution $N(0, \sigma^2)$. For the latter, the prior probability is $\pi_E = P(Y = E) = \frac{1}{1+\sqrt{2\pi}}$ and the class conditional has the exponential distribution.

$$P(X = x|Y = E) = \begin{cases}\lambda e^{-\lambda x} & \text{if } x \geq 0\\ 0 & \text{if } x < 0\end{cases}$$

Write an equation in x for the decision boundary. (Only the positive solutions of your equation will be relevant; ignore all $x < 0$.) Simplify the equation until it is quadratic in x. (You dont need to solve the quadratic equation. It should contain the constants σ and λ. Ignore the fact that 0 might or might not also be a point in the decision boundary.) (10 points)
Solution:

$$P(Y = N|X = x) = P(Y = E|X = x)$$
$$P(X = x|Y = N)P(Y = N) = P(X = x|Y = E)P(Y = E)$$
$$\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{x^2}{2\sigma^2})\frac{\sqrt{2\pi}}{1+\sqrt{2\pi}} = \lambda e^{-\lambda x}\frac{1}{1+\sqrt{2\pi}}$$
$$-\ln\sigma - \frac{x^2}{2\sigma^2} = \ln\lambda - \lambda x$$
$$\frac{x^2}{2\sigma^2} - \lambda x + \ln\sigma + \ln\lambda = 0.$$

6. Given data $\{(x_i, y_i) \in R^d \times \{0, 1\}\}_{i=1}^n$ and a query point x, we choose a parameter vector θ to minimize the loss (which is simply the negative log likelihood, weighted appropriately):

$$l(\theta; x) = -\sum_{i=1}^{n} w_i(x)[y_i \log(\mu(x_i)) + (1 - y_i)\log(1 - \mu(x_i))]$$

where

$$\mu(x_i) = \frac{1}{1+e^{-\theta \cdot x_i}}, w_i(x) = \exp(-\frac{||x-x_i||^2}{2\tau})$$

where τ s a hyperparameter that must be tuned. Note that whenever we receive a new query point x, we must solve the entire problem again with these new weights $w_i(x)$.

(a) Given a data point x, derive the gradient of $l(\theta; x)$ with respect to θ. (4 points)

$$\begin{aligned}\nabla_\theta l(\theta; x) &= -\sum_{i=1}^{n} w_i(x)(y_i - \mu(x_i))x_i \\ &= -\mathbf{X}^T z,\end{aligned}$$

where $z_i = w_i(x)(y_i - \mu(x_i))$.

(b) Given a data point x, derive the Hessian of $l(\theta; x)$ with respect to θ. (4 points)

$$\begin{aligned}H_\theta l(\theta; x) &= -\sum_{i=1}^{n} w_i(x)\mu(x_i)(1-\mu(x_i))x_i x_i^T \\ &= \mathbf{X}^T D\mathbf{X},\end{aligned}$$

where $D_{ii} = w_i(x)\mu(x_i)(1-\mu(x_i))$, $D_{ij} = 0$ if $i \neq j$.

(c) Given a data point x, write the update formula for Newtons method. (2 points)

$$\theta^{(t+1)} = \theta^{(t)} + [\mathbf{X}^T D\mathbf{X}]^{-1}\mathbf{X}^T z.$$

7. Now we discuss Bayesian inference in coin flipping. Let's denote the number of heads and the total number of trials by N_1 and N, respectively.

(a) Please derive the MAP estimation based on the prior $p(\theta) = \text{Beta}(\theta|\alpha, \beta)$. (4 points)
The Beta distribution is defined by

$$P(\theta) = \text{Beta}(\alpha, \beta) = \frac{\theta^{\beta-1}(1-\theta)^{\alpha-1}}{B(\alpha, \beta)}.$$

Combining with the expression for $P(D|\theta)$, we have:

$$\begin{aligned}\hat{\theta}^{MAP} &= \arg\max_\theta P(D|\theta)P(\theta) \\ &= \arg\max_\theta \theta^{N_1}(1-\theta)^{N_0}\frac{\theta^{\beta-1}(1-\theta)^{\alpha-1}}{B(\alpha, \beta)} \\ &= \arg\max_\theta \frac{\theta^{N_1+\beta-1}(1-\theta)^{N_0+\alpha-1}}{B(\alpha, \beta)} \\ &= \arg\max_\theta \theta^{N_1+\beta-1}(1-\theta)^{N_0+\alpha-1}.\end{aligned}$$

We calculate the derivative of the log of the likelihood function:

$$\begin{aligned}\frac{\partial \ell(\theta)}{\partial \theta} &= \frac{\partial \ln P(D|\theta)P(\theta)}{\partial \theta} \\ &= \frac{\partial \ln\left[\theta^{N_1+\beta-1}(1-\theta)^{N_0+\alpha-1}\right]}{\partial \theta} \\ &= \frac{\partial\left[N_1 \ln\theta + N_0 \ln(1-\theta)\right]}{\partial \theta} \\ &= (N_1+\beta_1-1)\frac{\partial \ln\theta}{\partial \theta} + (N_0+\alpha-1)\frac{\partial \ln(1-\theta)}{\partial \theta} \\ &= (N_1+\beta_1-1)\frac{\partial \ln\theta}{\partial \theta} + (N_0+\alpha-1)\frac{\partial \ln(1-\theta)}{\partial(1-\theta)}\cdot\frac{\partial(1-\theta)}{\partial \theta} \\ \frac{\partial \ell(\theta)}{\partial \theta} &= (N_1+\beta_1-1)\frac{1}{\theta} - (N_0+\alpha-1)\frac{1}{(1-\theta)}.\end{aligned}$$

Therefore,

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) = \frac{N_1 + \beta - 1}{N + \beta + \alpha - 2}.$$

(b) Please derive the MAP estimation based on the following prior:

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.4 \\ 0 & \text{otherwise,} \end{cases}$$

that believes the coin is fair, or is slightly biased towards tails. (4 points)
With the prior, the posterior becomes

$$P(D|\theta)P(\theta) = \begin{cases} 0.5 \cdot 0.5^{N_1}(1-0.5)^{N_0} & \theta = 0.5 \\ 0.5 \cdot 0.4^{N_1}(1-0.4)^{N_0} & \theta = 0.4 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 0.5^{N+1} & \theta = 0.5 \\ 0.5 \cdot 0.4^{N_1}0.6^{N-N_1} & \theta = 0.4 \\ 0 & \text{otherwise} \end{cases}$$

Since the value of θ only can be taken 0.5 or 0.4, we just need to compare two posteriors as follows:

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) = \begin{cases} 0.5 & \text{if } 0.5^{N+1} > 0.5 \cdot 0.4^{N_1}0.6^{N-N_1}, \\ 0.4 & \text{if } 0.5^{N+1} < 0.5 \cdot 0.4^{N_1}0.6^{N-N_1}. \end{cases}$$

Here, we don't consider the case of $0.5^{N+1} = 0.5 \cdot 0.4^{N_1}0.6^{N-N_1}$. After some simple computations, we have the solution:

$$\hat{\theta}^{MAP} = \begin{cases} 0.5 & \text{if } N < \frac{\ln 3 - \ln 2}{\ln 6 - \ln 5} N_1, \\ 0.4 & \text{if } N > \frac{\ln 3 - \ln 2}{\ln 6 - \ln 5} N_1. \end{cases}$$

(c) Suppose the true parameter is $\theta = 0.41$. Which prior leads to a better estimate when N is small? Which prior leads to a better estimate when N is large? (2 points)
When N is small, the prior in (b) leads a better estimate since the prior is a summary of our subjective beliefs about the data. When N is large, the estimate in (a) is better according to the law of large number.