ECE 271A: Homework #3

Due on November 13, 2023 at 11:59pm

Professor Vasconcelos

Ray Tsai

A16848188

Problem 1

In this problem we will consider the issue of linear regression and the connections between maximum likelihood and least squares solutions. Consider a problem where we have two random variables Z and X, such that

$$z = f(x, \theta) + \epsilon \tag{1}$$

where f is a polynomial with parameter vector θ

$$f(x, \theta) = \sum_{k=0}^{K} \theta_k x^k$$

and ϵ a Gaussian random variable of zero mean and variance σ^2. Our goal is to estimate the best estimate of the function given i.i.d. sample $\mathcal{D} = \{(\mathcal{D}_x, \mathcal{D}_z)\} = \{(x_1, z_1), \ldots, (x_n, z_n)\}$.

Part A

Formulate the problem as one of least squares, i.e define $z = (z_1, \ldots, z_n)^T$,

$$\Phi = \begin{bmatrix} 1 & \ldots & x_1^K \\ \vdots & \ddots & \vdots \\ 1 & \ldots & x_n^K \end{bmatrix}$$

and find the value of θ that minimizes

$$\|z - \Phi\theta\|^2.$$

Solution

We attempt to find θ, such that it gives a closest solution to

$$\Phi\theta = z.$$

By performing least squares, we get

$$\theta = (\Phi^T\Phi)^{-1}\Phi^T z.$$

Part B

Formulate the problem as one of ML estimation, i.e. write down the likelihood function $P_{Z|X}(z|x;\theta)$, and compute the ML estimate, i.e. the value of θ that maximizes $P_{Z|X}(\mathcal{D}_z|\mathcal{D}_x;\theta)$. Show that this is equivalent to part A.

Solution

Since X is known, $P_{Z|X}(z|x;\theta)$ becomes a Gaussian distribution with mean $f(x,\theta)$ and variance σ^2, namely

$$P_{Z|X}(z|x;\theta) = G(x, f(x,\theta), \sigma^2).$$

Given sampple $\mathcal{D}$, we take the natural log of $P_{Z|X}(\mathcal{D}_z|\mathcal{D}_x;\theta)$ and get

$$\begin{aligned}\theta^* &= \arg\max_\theta \sum_{i=1}^n -\frac{(z_i - f(x_i,\theta))^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) \\ &= \arg\min_\theta \sum_{i=1}^n (z_i - f(x_i,\theta))^2 \\ &= \arg\min_\theta \|z - \Phi\theta\|^2,\end{aligned}$$

and what we're looking for is obviously identical to the question in part A.

Part C

(The advantage of the statistical formulation is that makes the assumptions explicit. We will now challenge some of these.) Assume that instead of a fixed variance σ^2 we now have a variance that depends on the sample point, i.e.

$$z_i = f(x_i,\theta) + \epsilon_i,$$

where $\epsilon_i \sim N(0,\sigma_i^2)$. This means that our sample is independent but no longer identically distributed. It also means that we have different degrees of confidence in the different measurements (z_i, x_i). Redo part B under these conditions.

Solution

Instead of looking at individual data points, we view $\mathcal{D}_z$ and $\mathcal{D}_x$ as random vectors. Thus, the probablity distribution becomes a Gaussian distribution with mean $\Phi\theta$ and variance $\Sigma = \text{diag}(\sigma_1^2,\ldots,\sigma_n^2)$, namely

$$P_{Z|X}(z|\mathcal{D}_x;\theta) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp -\frac{1}{2}(z-\Phi\theta)^T\Sigma^{-1}(z-\Phi\theta).$$

We again do the log trick and get

$$\begin{aligned}\theta^* &= \arg\max_\theta \ln\left(\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\right) - \frac{1}{2}(z-\Phi\theta)^T\Sigma^{-1}(z-\Phi\theta) \\ &= \arg\min_\theta (z-\Phi\theta)^T\Sigma^{-1}(z-\Phi\theta).\end{aligned}$$

Let $g(\theta) = (z-\Phi\theta)^T\Sigma^{-1}(z-\Phi\theta)$. We take the gradient of g with respect to θ and get

$$\nabla_\theta g = -2\Phi^T\Sigma^{-1}(z-\Phi\theta) = 0.$$

Thus, we get a critical point $\theta^* = (\Phi^T\Sigma^{-1}\Phi)^{-1}\Phi^T\Sigma^{-1}z$. We take the Hessian of g and get that

$$\nabla_\theta^2 g = 2\Phi^T\Sigma^{-1}\Phi = 2(S\Phi)^T(S\Phi),$$

where $S = diag(\sigma_1^{-1},\ldots,\sigma_n^{-1})$. Since $\nabla_\theta^2 g$ can be decomposed into a product of a matrix and its transpose, it is positive definite, and so θ^* is the minimum point.

Part D

Consider the weighted least squares problem where the goal is to minimize

$$(z - \Phi\theta)^T W (z - \Phi\theta),$$

where W is a symmetrix matrix. Compute the optimal θ in this situation. What is the equivalent maximum likelihood problem? Rewrite the model (1), making explicit all the assumptions that lead to the new problem. What is the statistical interpretation of W?

Solution

By part C, we know the least square solution to this problem is

$$\theta^* = (\Phi^T W \Phi)^{-1} \Phi^T W z.$$

We can thus assume that $W = \Sigma^{-1}$ is the inverse of the covariance matrix, such that the random noise vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \sim N(0, \Sigma)$. Thus, (1) can be rewritten into

$$z = \Phi\theta + \epsilon,$$

where $z = (z_1, \ldots, z_n)^T$.

Part E

The L_2 norm is known to be prone to large estimation error if there are outliers in the training sample. These are training examples (z_i, x_i) for which, due to measurement errors or other extraneous causes, $|z_i - \sum_k \theta_i x_i^k|$ is much larger than for the remaining examples (the *inliers*). In fact, it is known that a single outlier can completely derail the least squares solution, an highly undesirable behavior. It is also well known that other norms lead to much more robust estimators. One of such distance metrics is the L_1-norm

$$L_1 = \sum_i \left| z_i - \sum_k \theta_k x_i^k \right|.$$

In the maximum likelihood framework, which is the statistical assumption that leads to the L_1 norm? Once again, rewrite the model (1), making explicit all the assumptions that lead to the new problem. Can you justify why this alternative formulation is more robust? In particular, provide a justification for i) why the L_1 norm is more robust to outliers, and ii) the associated statistical model (1) copes better with them.

Solution

We assume that the likelihood function is of the form $P_{Z|X}(z|x;\theta) = \alpha e^{-\frac{|z-f(x,\theta)|}{\sigma^2}}$, where α is a constant for normalization. Since $\int_{-\infty}^{\infty} e^{-\frac{|z-f(x,\theta)|}{\sigma^2}} dz = 2\int_0^{\infty} e^{-\frac{z-f(x,\theta)}{\sigma^2}} dz = 2\sigma^2$, we get $\alpha = \frac{1}{2\sigma^2}$. To examine the robustness of L_1 and L_2 norms, we compare the equations we attempt to minimize, namely $|z - f(x,\theta)|$ and $\|z - f(x,\theta)\|^2$. Consider a outlier z' such that its noise ϵ' is of great magnitude. Since when $|\epsilon'|$ is large, $\|z - f(x,\theta)\|^2 \gg |z' - f(x,\theta)|$, which implies that the L_2 norm penalize outliers a lot more then the L_1 norm. Therefore, the L_1 better accomodates the outliers compared to the L_2 norm, and thus the L_1 norm is more robust when the dataset has a few extreme outliers.

Problem 2

The purpose of this problem is to derive the Bayesian classifier for the d-dimensional multivariate Bernoulli case. As usual, work with each class separately, interpreting $P_{X|T}(x|\mathcal{D})$ to mean $P_{X|T_i,Y_i}(x|\mathcal{D}_i,\omega_i)$. Let the conditional probability for a given category be given by

$$P_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i},$$

and let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n samples independently drawn according to this probability density.

Part A

i. If $\mathbf{s} = (s_1, \ldots, s_d)^T$ is the sum of the n samples, show that

$$P_{T|\Theta}(\mathcal{D}|\theta) = \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i}.$$

Solution

$$\begin{aligned} P_{T|\Theta}(\mathcal{D}|\theta) &= \prod_{j=1}^{n} P_{\mathbf{X}_j|\Theta}(\mathbf{x}_j|\theta) \\ &= \prod_{j=1}^{n}\prod_{i=1}^{d} \theta_i^{x_{ji}}(1-\theta_i)^{1-x_{ji}} \\ &= \prod_{i=1}^{d} \theta_i^{\sum_{j=1}^{n} x_{ji}}(1-\theta_i)^{\sum_{j=1}^{n} 1-x_{ji}} \\ &= \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i}. \end{aligned}$$

ii. Assuming a uniform a priori distribution for θ and using the identity

$$\int_0^1 \theta^m(1-\theta)^n d\theta = \frac{m!n!}{(m+n+1)!},$$

show that

$$P_{\Theta|T}(\theta|\mathcal{D}) = \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!}\theta_i^{s_i}(1-\theta_i)^{n-s_i}.$$

Solution

$$\begin{aligned} P_{\Theta|T}(\theta|\mathcal{D}) &\propto P_{T|\Theta}(\mathcal{D}|\theta)P_{\Theta}(\theta) \\ &\propto \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i} \end{aligned}$$

Since

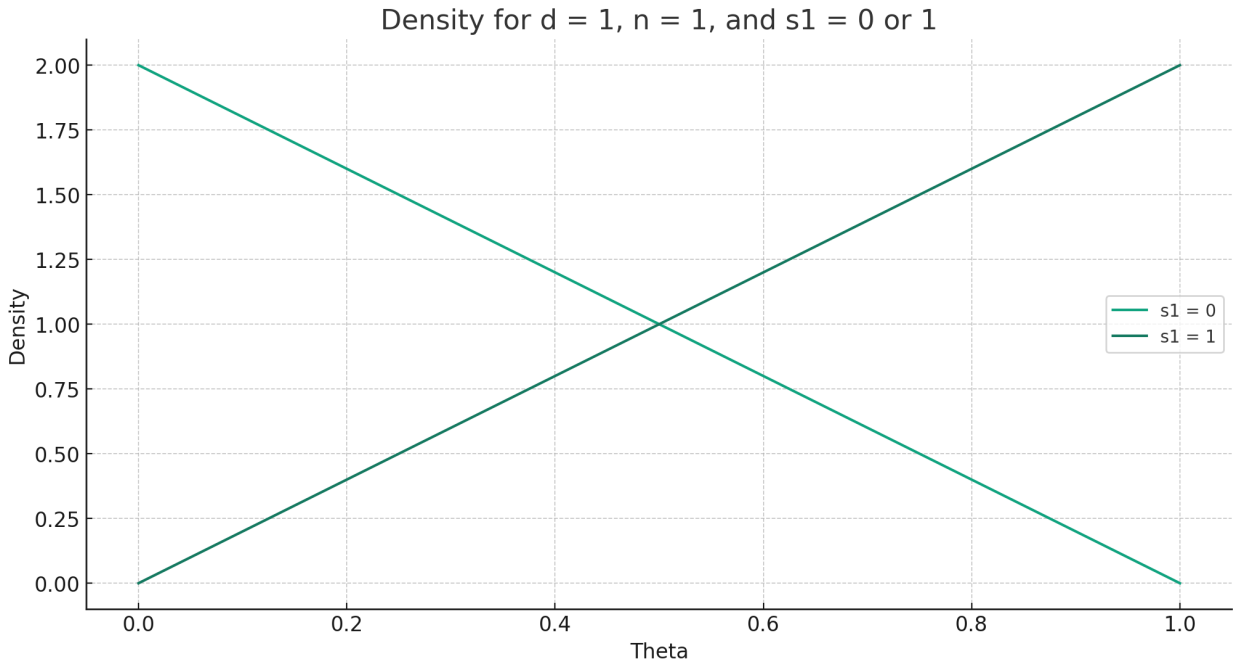
$$\int_0^1 \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i} d\theta = \prod_{i=1}^{d} \int_0^1 \theta_i^{s_i}(1-\theta_i)^{n-s_i} d\theta_i = \prod_{i=1}^{d} \frac{s_i!(n-s_i)!}{(n+1)!},$$

we have

$$P_{\Theta|T}(\theta|\mathcal{D}) = \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!}\theta_i^{s_i}(1-\theta_i)^{n-s_i}.$$

iii. Plot this density for the case $d = 1, n = 1$, and for the two resulting possibilities for s_1.

Solution When $d = 1, n = 1$, we have $P_{\Theta|T}(\theta|\mathcal{D}) = 2\theta_1^{s_1}(1-\theta_1)^{1-s_1}$, where $s_1 \in \{0, 1\}$. The plot for each cases of s_1 is shown below:



Note that an observation of 0 turns the uniform distribution into a distribution that weighs a lot more on smaller θ, and vice versa.

iv. Integrate the product $P_{X|\Theta}(x|\theta)P_{\Theta|T}(\theta|\mathcal{D})$ over θ to obtain the desired conditional probability.

Solution

$$
\begin{aligned}
P_{X|T}(\mathbf{x}|D) &= \int_0^1 P_{X|\Theta}(x|\theta)P_{\Theta|T}(\theta|\mathcal{D})d\theta \\
&= \int_0^1 \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i} \cdot \frac{(n+1)!}{s_i!(n-s_i)!}\theta_i^{s_i}(1-\theta_i)^{n-s_i} \\
&= \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!} \int_0^1 \theta_i^{x_i+s_i}(1-\theta_i)^{n-x_i-s_i+1} \\
&= \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!} \cdot \frac{(x_i+s_i)!(n-x_i-s_i+1)!}{(n+2)!} \\
&= \prod_{i=1}^{d} \frac{(s_i+1)^{x_i}(n-s_i+1)^{1-x_i}}{n+2} \\
&= \prod_{i=1}^{d} \left(\frac{s_i+1}{n+2}\right)^{x_i} \left(1-\frac{s_i+1}{n+2}\right)^{1-x_i}.
\end{aligned}
$$

v. If we think of obtaining $P_{X|T}(\mathbf{x}|\mathcal{D})$ by substituting an estimate $\hat{\theta}$ for θ in $P_{X|\Theta}(x|\theta)$, what is the effective Bayesian estimate for θ.

Solution

The effective estimate for θ is

$$\hat{\theta_i} = \frac{s_i + 1}{n + 2},$$

for $i \in \{1, \ldots, d\}$.

Part B

What is the ML estimate for θ in this problem? What is the MAP estimate for θ in this problem? Do you see any advantage in favoring one of the estimates in favor of the others? How does that relate to the uniform prior that was assumed for θ?

Solution

The ML estimate is

$$\begin{aligned}
\theta_{ML} &= \arg\max_{\theta} P_{T|\Theta}(\mathcal{D}|\theta) \\
&= \arg\max_{\theta} \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i} \\
&= \arg\max_{\theta} \sum_{i=1}^{d} s_i \ln\theta_i + (n-s_i)\ln(1-\theta_i).
\end{aligned}$$

Let $g(\theta) = \sum_{i=1}^{d} s_i \ln\theta_i + (n-s_i)\ln(1-\theta_i)$. Since $\frac{\partial g}{\partial \theta_i} = \frac{s_i}{\theta_i} - \frac{n-s_i}{1-\theta_i}$ and $\frac{\partial^2 g_i}{\partial \theta_i^2} = -\frac{s_i}{\theta_i^2} - \frac{n-s_i}{(1-\theta_i)^2} < 0$, we get $\theta_{ML} = \frac{1}{n}\mathbf{s}$.

The MAP estimate is

$$\begin{aligned}
\theta_{MAP} &= \arg\max_{\theta} P_{T|\Theta}(\mathcal{D}|\theta)P_{\Theta}(\theta) && (P_{\Theta}(\theta) \text{ is uniform}) \\
&= \arg\max_{\theta} P_{T|\Theta}(\mathcal{D}|\theta) \\
&= \theta_{ML}.
\end{aligned}$$

We can see that our assumption of the uniform prior made the two estimates identical, and thus there are no advantage favoring one over another in this case.

Problem 3

Consider problem 3 of the previous assignment, i.e. a random variable X such that $P_X(k) = \pi_k, k \in 1, \ldots, N$, n independent observations from X, a random vector $C = (C_1, \ldots, C_N)^T$ where C_k is the number of times that the observed value is k (i.e. C is the histogram of the sample of observations). We have seen that, C has multinomial distribution

$$P_{C_1,\ldots,C_N}(c_1, \ldots, c_N) = \frac{n!}{\prod_{k=1}^N c_k!} \prod_{j=1}^N \pi_j^{c_j}.$$

In this problem we are going to compute MAP estimates for this model. Notice that the parameters are probabilities and, therefore, not every prior will be acceptable here (since $\pi_j > 0$ and $\sum_j \pi_j = 1$ for the prior to be valid). One distribution over vectors $\pi = (\pi_1, \ldots, \pi_N)^T$ that satisfies this constraint is the Dirichlet distribution

$$P_{\Pi_1,\ldots,\Pi_N}(\pi_1, \ldots, \pi_N) = \frac{\Gamma\left(\sum_{j=1}^N u_j\right)}{\prod_{k=1}^N \Gamma(u_j)} \prod_{j=1}^N \pi_j^{u_j - 1},$$

where the u_j are the set of *hyperparameters* (parameters of the prior) and

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

the Gamma function.

Part A

Derive the MAP estimator for the parameters $\pi_i, i = 1, \ldots, N$ using the Dirichlet prior.

Solution

Let $\pi = (\pi_1, \ldots, \pi_N)$.

$$\begin{aligned}
\pi^* &= \arg\max_\pi P_{C_1,\ldots,C_N|\Pi}(c_1, \ldots, c_N|\pi) P_{\Pi_1,\ldots,\Pi_N}(\pi_1, \ldots, \pi_N) \\
&= \arg\max_\pi \frac{n!\Gamma\left(\sum_{j=1}^N u_j\right)}{\prod_{k=1}^N c_k!\Gamma(u_j)} \prod_{j=1}^N \pi_j^{c_j + u_j - 1} \\
&= \arg\max_\pi -\sum_{j=1}^N (c_j + u_j - 1) \ln \pi_j.
\end{aligned}$$

Let $g(\pi) = -\sum_{j=1}^N (c_j + u_j - 1) \ln \pi_j - \lambda\left(\sum_{j=1}^N \pi_j - 1\right)$. Since $\frac{\partial g}{\partial \pi_j} = \frac{c_j + u_j - 1}{\pi_j} - \lambda$ and $\frac{\partial^2 g}{\partial \pi_j^2} = -\frac{c_j + u_j - 1}{\pi_j^2} < 0$, we know $c_j + u_j - 1 = \lambda \pi_j$. Summing over j, we get $\lambda = n - N + \sum_j u_j$, and so $\pi_j^* = \frac{c_j + u_j - 1}{\lambda} = \frac{c_j + u_j - 1}{n + \sum_j u_j - N}$.

Part B

Compare this estimator with the ML estimator derived in the previous assignment. What is the use of this prior equivalent to, in terms of the ML solution?

Solution

Since $\lim_{n\to\infty} \pi_j^* = \frac{c_j}{n+\sum_j u_j - N}$, we know $\lim_{n\to\infty} \frac{\pi_j^*}{\pi_{MAP\,j}} = \frac{n+\sum_j u_j - N}{n} = 1$, and thus π^* approaches the ML estimate when n grows large. Conversely, when $n = 0$, $\pi_j^* = \frac{u_j - 1}{\sum_j u_j - N}$. Thus, we can interpret c_j and n as the observed counts and trails, and view $u_j - 1$ and $\sum_j u_j - N$ as the supposed counts and trails from our knowledge.

Part C

What is the effect of the prior as the number of samples n increases? Does this make intuitive sense?

Solution

From part B, we see that the MAP estimate appraoches the ML estimate when n grows large. This means that the prior becomes a lot less influencial compared to our estimation when the sample size is large, which aligns with the idea of Bayesian estimation.

Part D

In this problem and problem 2 we have seen two ways of avoiding the computational complexity of computing a fully Bayesian solution: i) to rely on a non-informative prior, and ii) to rely on an informative prior and compute the MAP solution. Qualitatively, what do you have to say about the results obtained with the two solutions? What does this tell you about the robustness of the Bayesian framework?

Solution

Regardless of the prior, the Bayesian framework is qualitatively equivalent to the ML estimate when n is large, which implies that the Bayesian framework is really robust.