UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: PROBABILITY AND RANDOM PROCESSES

Homework 11
Spring 2023

1. **Estimating Parameter of Random Graph Given Average Degree**

   Consider an Erdős–Rényi random graph on n vertices, in which each edge appears independently with probability p. Let D be the average degree of a vertex in the graph. Compute the maximum likelihood estimator of p given D. You may approximate $\text{Binomial}(k, p) \approx \text{Poisson}(kp)$.

   Solution: Let m be the number of edges in the graph, so that $D = \frac{2m}{n}$ by the handshake lemma. Write $M = \binom{n}{2}$. Since m has distribution $\text{Binomial}(M, p) \approx \text{Poisson}(Mp)$,

$$\mathbb{P}(D = d; p) \approx \frac{M^{nd/2}}{(nd/2)!} p^{nd/2} e^{-Mp}.$$

   To obtain the log-likelihood, we take the logarithm and drop all terms which have no dependence on p, which gives the function

$$\ell(d; p) \approx -\binom{n}{2} p + \frac{nd}{2} \ln p.$$

   Differentiating w.r.t. p, we see that the MLE for p is $\hat{p} = \frac{D}{n-1}$, which agrees with intuition: the average degree of a node is Binomial with $n - 1$ potential neighbors and probability p for each edge, so the expected value of D is $(n - 1)p$.

2. **Community Detection Using MAP**

It may be helpful to work on this problem in conjunction with the relevant lab. The *stochastic block model* (SBM) defines the random graph $\mathcal{G}(n, p, q)$ consisting of two communities of size $\frac{n}{2}$ each, such that the probability an edge exists between two nodes of the same community is p, and the probability an edge exists between two nodes in different communities is $q < p$. The goal of the problem is to exactly determine the two communities, given only the graph.

Show that the MAP estimate of the two communities is equivalent to finding the *min-bisection* or *balanced min-cut* of the graph, the split of G into two groups of size $\frac{n}{2}$ that has the minimum edge weight across the partition. Assume that any assignment of the communities is a priori equally likely.

Solution: Let $G \sim \mathcal{G}(n, p, q)$, and let A be a random variable representing the assignment or labelling of the two communities. We are interested in

$$\mathrm{MAP}(A \mid G) = \operatorname*{argmax}_A \mathbb{P}(G \mid A) \cdot \mathbb{P}(A) = \operatorname*{argmax}_A \mathbb{P}(G \mid A).$$

Note that the MAP rule is equivalent to the MLE as each assignment of labels is equally likely. Let k be the number of edges across the partition in assignment A, and let m be the number of edges in G. Then

$$\begin{aligned}\mathbb{P}(G \mid A) &= q^k (1-q)^{(\frac{n}{2})^2 - k} \cdot p^{m-k} (1-p)^{2\binom{n/2}{2} - (m-k)} \\ &= \left(\frac{q}{1-q} \cdot \frac{1-p}{p}\right)^k \cdot \left(\frac{p}{1-p}\right)^m \cdot (1-p)^{2\binom{n/2}{2}} \cdot (1-q)^{n^2/4}.\end{aligned}$$

Now, the last three terms do not depend on the assignment of labels, and thus do not affect the likelihood function. We also see that

$$p > q \implies \left(\frac{q}{1-q} \cdot \frac{1-p}{p}\right) < 1,$$

so increasing k corresponds to decreasing the likelihood. Therefore, the MAP rule is to select the partition with the smallest number of edges across it, which is exactly the min-bisection of the graph.

3. **MLE of Uniform Distribution**

Find the MLE of θ given $X_1, \ldots, X_n \sim_{\text{i.i.d.}} \text{Uniform}([0, \theta])$.

Solution: The likelihood function is given by

$$L(\theta \mid \mathbf{x}) = L(\theta \mid x_1, \ldots, x_n) = \frac{1}{\theta^n} \mathbb{1}_{0 \le x_1, \ldots, x_n \le \theta} .$$

Taking the derivative of the log-likelihood, we have

$$\frac{\partial \ell(\theta \mid \mathbf{x})}{\partial \theta} = -\frac{n}{\theta} \mathbb{1}_{0 \le x_1, \ldots, x_n \le \theta},$$

so the likelihood is decreasing in θ for $\theta \ge X_1, \ldots, X_n$. Hence, the likelihood is maximized at $\theta^* = X_{(n)} = \max\{X_1, \ldots, X_n\}$.

4. **Linear Regression, MLE, and MAP**

Suppose you draw n i.i.d. data points $(x_1, y_1), \dots, (x_n, y_n)$, where the true relationship is given by $Y = WX + \varepsilon$ for $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. In other words, Y has a linear dependence on X with additive Gaussian noise.

a. Show that finding the MLE of W given the data points $\{(x_i, y_i)\}_{i=1}^n$ is equivalent to minimizing mean squared error, or minimizing the cost function

$$J(w) = \sum_{i=1}^{n} (y_i - wx_i)^2.$$

b. Now suppose that W has a *Laplace* prior distribution,

$$f_W(w) = \frac{1}{2\beta} e^{-|w|/\beta}.$$

Show that finding the MAP estimate of W given the data points $\{(x_i, y_i)\}_{i=1}^n$ is equivalent to minimizing the cost function

$$J(w) = \sum_{i=1}^{n} (y_i - wx_i)^2 + \lambda |w|.$$

(You should determine what λ is.) This is interpreted as a one-dimensional ℓ^1-regularized least-squares criterion, also known as LASSO.

Solution:

a. The likelihood of the data is

$$L((x_1, y_1), \dots, (x_n, y_n) \mid W = w) = \prod_{i=1}^{n} L((x_i, y_i) \mid W = w)$$

as the data points are conditionally independent given W;

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - wx_i)^2/(2\sigma^2)}$$

as the likelihood of (x_i, y_i) given $W = w$ is the density of ε_i evaluated at $y_i - wx_i$;

$$\propto \prod_{i=1}^{n} e^{-(y_i - wx_i)^2/(2\sigma^2)},$$

discarding constant factors that do not depend on the data points or w. We now find it more convenient to work with the log-likelihood

$$\ell((x_1, y_1), \dots, (x_n, y_n) \mid W = w) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - wx_i)^2.$$

We wish to maximize the log-likelihood with respect to w, which is equivalent to *minimizing* the cost function

$$J(w) = \sum_{i=1}^{n} (y_i - wx_i)^2.$$

b. The likelihood of W given the data points is

$$\begin{aligned}
L(w \mid (x_1,y_1),\ldots,(x_n,y_n)) &\propto L((x_1,y_1),\ldots,(x_n,y_n) \mid W=w)\cdot f_W(w) \\
&= f_W(w)\prod_{i=1}^{n} L((x_i,y_i) \mid W=w) \\
&= \frac{1}{2\beta}e^{-|w|/\beta}\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-(y_i-wx_i)^2/(2\sigma^2)} \\
&\propto e^{-|w|/\beta}\prod_{i=1}^{n} e^{-(y_i-wx_i)^2/(2\sigma^2)}.
\end{aligned}$$

Again, we find it more convenient to work with the log-likelihood

$$\ell(w \mid (x_1,y_1),\ldots,(x_n,y_n)) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-wx_i)^2 - \frac{1}{\beta}|w|.$$

Maximizing the log-likelihood is equivalent to minimizing the cost function

$$J(w) = \sum_{i=1}^{n}(y_i-wx_i)^2 + \lambda|w|$$

with $\lambda = 2\sigma^2/\beta$.

5. **Poisson Process MAP**

Customers arrive to a store according to a Poisson process with rate 1. The store manager learns of a rumor that one of the employees is sending every other customer to the rival store, so that *deterministically*, every odd-numbered customer $1, 3, 5, \ldots$ is sent away.

Let $X = 1$ be the hypothesis that the rumor is true and $X = 0$ the rumor is false, assuming that both hypotheses are equally likely. Suppose a customer arrives to the store at time 0. After that, the manager observes $T_1, \ldots, T_n$, where T_i is the time of the ith subsequent sale, $i = 1, \ldots, n$. Derive the MAP rule to determine whether the rumor was true or not.

Solution: Note that both hypotheses are a priori equally likely, so the MAP rule is equivalent to the MLE rule. Also note that the interarrival times τ_i are independent whether conditioned on $X = 1$ or on $X = 0$. The density of an interarrival interval given $X = 1$ is Erlang of order 2, so for $0 \le t_1 < \cdots < t_n$,

$$f_{T_1,\ldots,T_n|X}(t_1, \ldots, t_n \mid 1) = \prod_{i=1}^{n} (t_i - t_{i-1}) e^{-(t_i - t_{i-1})} = e^{-t_n} \prod_{i=1}^{n} (t_i - t_{i-1}).$$

The density of an interarrival interval given $X = 0$ is Exponential, so

$$f_{T_1,\ldots,T_n|X}(t_1, \ldots, t_n \mid 0) = e^{-t_n}.$$

Taking the logarithm of both expressions, we see that the MAP is to declare $X = 1$ whenever

$$\sum_{i=1}^{n} \ln(T_i - T_{i-1}) \ge 0.$$

6. **Minimum-Error Property of MAP**

a. Let $X \in \{0,1\}$, and suppose we have the prior $\mathbb{P}(X=0)=\pi_0$ and $\mathbb{P}(X=1)=\pi_1$. Let $\hat{X}_{\text{MAP}}$ be the MAP estimate of X given the random variable Y, and let $\hat{X}$ be any other estimate of X given Y. Show that

$$\mathbb{P}(X \neq \hat{X}_{\text{MAP}}) \leq \mathbb{P}(X \neq \hat{X}).$$

b. Now, also suppose that type I errors (declaring $\hat{X}=1$ when $X=0$) incur a cost of $c_1 \geq 0$ and type II errors (declaring $\hat{X}=0$ when $X=1$) a cost of $c_2 \geq 0$. Derive the decision rule $\hat{X}$ that minimizes the total cost

$$c_1\,\mathbb{P}(\hat{X}=1, X=0) + c_2\,\mathbb{P}(\hat{X}=0, X=1).$$

Solution:

a. We write $\hat{X}_{\text{MAP}} = r^*(Y)$, where

$$r^*(y) = \operatorname*{argmax}_x \mathbb{P}(X=x, Y=y) = \operatorname*{argmin}_x \mathbb{P}(X \neq x, Y=y).$$

Now, the error probability for a general estimate $\hat{X}$ is

$$\begin{aligned}
\mathbb{P}(X \neq \hat{X}) &= \sum_y \mathbb{P}(X \neq \hat{X}, Y=y) \\
&= \sum_y \sum_z \mathbb{P}(X \neq z, Y=y) \cdot \mathbb{P}(\hat{X}=z \mid Y=y) \\
&\geq \sum_y \sum_z \mathbb{P}(X \neq r^*(y), Y=y) \cdot \mathbb{P}(\hat{X}=z \mid Y=y) \\
&= \sum_y \mathbb{P}(X \neq r^*(y), Y=y) \\
&= \mathbb{P}(X \neq r^*(Y)).
\end{aligned}$$

Remark. $\hat{X}$ being an estimate of X given Y means that it is conditionally independent of X given Y; that is, $X \to Y \to \hat{X}$ forms a Markov chain, as we saw in HW 06 Q4 and HW 07 Q1. This allowed us to drop the conditioning on X in the term $\mathbb{P}(\hat{X}=z \mid Y=y)$.

Remark. The error probability $\mathbb{P}(X \neq \hat{X}) = \mathbb{E}(\mathbb{1}\{X \neq \hat{X}\})$ is also known as the *Bayes risk* of $\hat{X}$ under the 0–1 loss function. We have shown that $\hat{X}_{\text{MAP}}$ minimizes $\mathbb{E}(\mathbb{1}\{X \neq \hat{X}\})$, i.e. MAP is the *Bayes-optimal* decision rule for estimating $X \in \{0,1\}$ under 0–1 loss.

Alternate solution. As $X \in \{0,1\}$, the MAP estimate is the threshold decision rule

$$\hat{X}_{\text{MAP}} = \mathbb{1}\{p_{Y|X}(Y \mid 1)\cdot \pi_1 \geq p_{Y|X}(Y \mid 0)\cdot \pi_0\} = \mathbb{1}\{L(Y) \geq \tfrac{\pi_0}{\pi_1}\},$$

$\pi_1 > 0$ without loss of generality. We can rewrite the error probability for $\hat{X}$ as

$$\begin{aligned}
\mathbb{P}(X \neq \hat{X}) &= \pi_0\,\mathbb{P}(\hat{X}=1 \mid X=0) + \pi_1\,\mathbb{P}(\hat{X}=0 \mid X=1) \\
&= \pi_0\,\mathbb{E}(\hat{X} \mid X=0) + \pi_1(1 - \mathbb{E}(\hat{X} \mid X=1))
\end{aligned}$$

$$= \pi_1 \, \mathbb{E}(\tfrac{\pi_0}{\pi_1} \hat{X} \mid X = 0) + \pi_1 - \pi_1 \, \mathbb{E}(L(Y) \hat{X} \mid X = 0)$$

$$= \pi_1 - \pi_1 \, \mathbb{E}((L(Y) - \tfrac{\pi_0}{\pi_1}) \hat{X} \mid X = 0).$$

Observe that $(L(Y) - \frac{\pi_0}{\pi_1}) \hat{X}_{\mathrm{MAP}} \geq (L(Y) - \frac{\pi_0}{\pi_1}) \hat{X}$ by the definition of $\hat{X}_{\mathrm{MAP}}$. Thus the error probability of the MAP estimate is minimal.

b. Suppose $c_1 + c_2 > 0$ without loss of generality, and let $c \coloneqq c_1 \pi_0 + c_2 \pi_1$. The total cost of $\hat{X}$ is precisely

$$c_1 \pi_0 \, \mathbb{P}(\hat{X} = 1 \mid X = 0) + c_2 \pi_1 \, \mathbb{P}(\hat{X} = 0 \mid X = 1) = c \, \mathbb{P}(X \neq \hat{X}),$$

c times the error probability of $\hat{X}$ for the prior $\mathbb{P}(X = 0) = \frac{c_1 \pi_0}{c}$ and $\mathbb{P}(X = 1) = \frac{c_2 \pi_1}{c}$. By part a, the total cost is minimized by the MAP estimate under this reweighted prior.