CS 475 Machine Learning: Homework 1 Analytical (35 points)

Assigned: Monday, September 13, 2021

Due: Wednesday, September 22, 2021, 11:59 pm US/Eastern

Partner 1: Dimitri Lezcano (dlezcan1), Partner 2: Harrison Khoo (hkhoo2)

Wednesday, September 22, 2021

Instructions

We have provided this LaTeX document for turning in this homework. We give you one or more boxes to answer each question. The question to answer for each box will be noted in the title of the box. You can change the size of the box if you need more space.

Other than your name, do not type anything outside the boxes. Leave the rest of the document unchanged.

Do not change any formatting in this document, or we may be unable to grade your work. This includes, but is not limited to, the height of textboxes, font sizes, and the spacing of text and tables. Additionally, do not add text outside of the answer boxes. Entering your answers are the only changes allowed.

We strongly recommend you review your answers in the generated PDF to ensure they appear correct. We will grade what appears in the answer boxes in the submitted PDF, NOT the original latex file.

1 Probability and Linear Algebra: Diagnostic

This section is ungraded and intended for diagnostic purposes only. While answers to these questions are easy to compute with access to a statistical language interpreter, or look up on the internet, we advise you not to do so. These questions are an opportunity to verify that you feel comfortable with the prerequisite topics for this class. If you don't know/remember everything, that doesn't mean you can't still do well, but you would need to put in extra effort reviewing the relevant background.

Probability

1. Recall that variance is defined as $\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$. Prove that $\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$.

2. Let X be a random variable such that $X = YZ$, where $Y \sim \mathcal{N}(0, \sigma^2)$ and $Z \sim \mathrm{Bernoulli}(p)$. Find the mean and variance of X.

Linear Algebra

1. Show that the vector w is orthogonal to the hyperplane $w^T x + b = 0$.

2. Consider the matrix A below:

$$A = \begin{bmatrix} 1 & 2 & 5 \\ 2 & 4 & 3 \\ 4 & 5 & 8 \end{bmatrix}$$

   (a) What is the rank of A?

   (b) Compute the determinant of A.

2 Likelihood

Given n data points $\{x_1, x_2, \dots, x_n\}$ and the following linear model

$$y_i = \omega^T x_i + \epsilon_i$$

where ϵ_i is a random variable representing the noise and is independent of $\mathbf{x}$.

(a) Assume ϵ_i comes from a standard Gaussian distribution, i.e.

$$P(\epsilon_i) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\epsilon_i^2}{2}\right)$$

Compute the conditional log-likelihood of $\mathbf{y}$ given $\mathbf{x}$ and ω. Give the simplest function of y_i and ω such that minimizing this function is equivalent to maximizing the conditional log-likelihood.

$$\begin{aligned}\mathcal{L}_{[D]}(\mathbf{y}|\mathbf{x}, w) &= \sum_{i=1}^{n} \ln p(y_i|x_i, w) = \sum_{i=1}^{n} \ln(\frac{1}{\sqrt{2\pi}} \exp\{\frac{-1}{2}(x_i - w^T x_i)^2\}) \\ &= \sum_{i=1}^{n} (\ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2}(y_i - w^T x_i)^2) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + n \ln \frac{1}{\sqrt{2\pi}}\end{aligned}$$

The constant term $n \ln \frac{1}{\sqrt{2\pi}}$ does not affect the minimization, so

$$\begin{aligned}argmax \ -\frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + n \ln \frac{1}{\sqrt{2\pi}} &= argmax \ -\frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 \\ &= argmin \frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2\end{aligned}$$

We can minimize the least square loss function, $\mathcal{L}(\mathbf{y}, w) = \sum_{i=1}^{n}(y_i - w^T x_i)^2$, to maximize the conditional log-likelihood.

(b) Assume ϵ_i comes from a Laplace distribution, i.e.

$$P(\epsilon_i) = \frac{1}{2} \exp\Big(- |\epsilon_i|\Big)$$

Compute the conditional log-likelihood of $\mathbf{y}$ given $\mathbf{x}$ and ω. Give the simplest function of y_i and ω such that minimizing this function is equivalent to maximizing the conditional log-likelihood

$$\begin{aligned}\mathcal{L}_{[D]}(\mathbf{y}|\mathbf{x}, w) &= \sum_{i=1}^{n} \ln p(y_i|x_i, \omega) = \sum_{i=1}^{n} \ln P(y_i - \omega^T x_i) \\ &= \sum_{i=1}^{n} \big(-|y_i - \omega^T x_i| - \ln 2\big) = -\sum_{i=1}^{n} \big(|y_i - \omega^T x_i|\big) - n \ln 2\end{aligned}$$

Since we have that $\arg\max \mathcal{L}_{[D]}$ is independent of constants, we have $\arg\max \mathcal{L}_{[D]} = \arg\max \mathcal{L}_{[D]} + n \ln 2 = \arg\min -(\mathcal{L}_{[D]} + n \ln 2)$. Thus, we can use a loss function of

$$\mathcal{L}(\mathbf{y}, \omega) = \sum_{i=1}^{n} |y_i - \omega^T x_i|$$

where minimizing the above is equivalent to maximizing the conditional log-likelihood.

(c) Which loss is easier to minimize? Which loss is more robust to outliers? Explain in detail.

(a) is easier to minimize because derivatives are well-defined $\forall \mathbf{y}$ and ω while the loss in (b) has is not differentiable everywhere (due to the absolute value). (b) is more robust to outliers since it weights less the discrepancy between y_i and $\omega^T x_i$ using an L1 norm while (a) uses an L2 norm grows with a squared relationship with the discrepancy between y_i and $\omega^T x_i$. We can see this if we assume that $|y_i - \omega^T x_i| = 2$ for some y_i, ω, x_i and for all other examples x_j, $j \neq i$, we have $|y_j - \omega^T x_j| = 0$. In (a), the contribution to the loss function would be 4 while in (b), the contribution to the loss function would be 2, resulting in total losses of 2 and 4, respectively. We can clearly see that outliers in (b) do not contribute as heavily to the loss function as in (a).

3 Conditional Independence

A large group of people were surveyed on their recent health. Of these, 0.20 had a fever and 0.05 had pneumonia. Among the people who had pneumonia, 0.70 had cough as a symptom and 0.50 had fever as a symptom. Among the people who had a fever, 0.40 had cough as a symptom.

Let us create a probabilistic model where the presence/absence of each of these two symptoms, cough and fever, are conditionally independent given the presence/absence of pneumonia. Using this data for the empirical probabilities of our model, answer the following questions.

1. Find the probability that someone has both a cough and a fever.

$$\begin{aligned} p(fever) &= p(f) = 0.2 \\ p(pneumonia) &= p(p) = 0.05 \\ p(cough|pneumonia) &= p(c|p) = 0.7 \\ p(fever|pneumonia) &= p(f|p) = 0.5 \\ p(cough|fever) &= p(c|f) = 0.4 \end{aligned}$$

$$P(c \cap f) = p(c|f) \times p(f) = 0.4 \times 0.2 = 0.08$$

2. Find the probability that someone has pneumonia given that they have a fever but no cough.

$$\begin{aligned} p(p|f \cap not\ c) &= \frac{p(f \cap not\ c|p) \times p(p)}{p(f \cap not\ c)} \\ &= \frac{p(f|p) \times p(not\ c|p) \times p(p)}{p(f) - p(f \cap c)} \\ &= \frac{p(f|p) \times (1 - p(c|p)) \times p(p)}{p(f) - p(f \cap c)} \\ &= \frac{(0.5) \times (1 - 0.7) \times (0.05)}{(0.2 - 0.08)} = \frac{0.0075}{0.12} = 0.0625 \end{aligned}$$

3. Given assumptions described above, how many parameters do we need to specify the joint distribution $p(fever, cough, pneumonia)$?

Given the assumption that $cough \perp\!\!\!\perp fever \mid pneumonia$, we have that

$$\begin{aligned} p(f, c, p) &= p(f, c \mid p)p(p) \\ &= p(f \mid p)p(c \mid p)p(p) \because (cough \perp\!\!\!\perp fever \mid pneumonia) \end{aligned}$$

Therefore, we only need 3 parameters $p(p), p(c \mid p)$, and $p(f \mid p)$ to fully specify the joint distribution $p(fever, cough, pneumonia)$.

4 Conjugate Priors

1. Define what a conjugate prior is.

The conjugate prior is a special case prior where the posterior for a given likelihood function is of the same distribution family as the prior.

2. Why are conjugate priors useful?

Conjugate priors may save computing time by reducing the required number and complexity of calculations. If the distribution is not closed-form, then we would have to otherwise rely on using an expensive optimization method, like gradient descent, to calculate the posterior.

3. Show that the Gamma distribution is a conjugate prior of the exponential distribution. That is, show that if $x \sim \text{Exp}(\lambda)$ and $\lambda \sim \text{Gamma}(\alpha, \beta)$, then $p(\lambda|x) \sim \text{Gamma}(\alpha^*, \beta^*)$ for some α^*, β^*.

$$\begin{aligned} p(\lambda|x) &\propto p(x|\lambda)p(\lambda) \\ &\propto \lambda^n e^{-\lambda \sum_i x_i} \lambda^{\alpha-1} e^{-\beta\lambda} \\ &\propto e^{-\lambda(\sum_i x_i - \beta)} \lambda^{n+\alpha-1} \\ p(\lambda|x) &\propto Gamma(\alpha^*, \beta^*), where\ \alpha^* = \alpha + n\ and\ \beta^* = \sum_i x_i + \beta \end{aligned}$$

5 Gibbs Sampling and the Semi-Graphoid Axioms

1. Assume a joint distribution $p(x_1, \dots, x_k)$ over binary random variables $X_1, \dots, X_k$. What's the size of the joint probability table?

For binary random variables, there are 2 options per variable and k total random variables. The total number of options that the joint distribution can take are 2^k. However, the restriction that $\sum_{\{x_i \mid i \in \{1,\dots,k\}\}} p(x_1, \dots, x_k) = 1$ is a constraint such that we can deduce one of the outcomes from the other distributions, i.e $p(x_1, \dots, x_k = 0) = 1 - \sum p(x_1, \dots, x_k = 1)$. Therefore the size of the joint probability table is $\mathbf{2^k} - \mathbf{1}$.

2. Assume $(X_1 \perp\!\!\!\perp X_3, \dots, X_k \mid X_2)$, $(X_k \perp\!\!\!\perp X_1, \dots, X_{k-2} \mid X_{k-1})$, and $(X_i \perp\!\!\!\perp X_1, \dots, X_{i-2}, X_{i+2}, \dots, X_k \mid X_{i-1}, X_{i+1})$ for each $i = 2, \dots, k-1$. What's the smallest number of parameters we would need to specify to create a Gibbs sampler for $p(x_1, \dots, x_k)$?

We have the following $\forall i \in [2, k-1]$ from the assumptions:

$$\begin{aligned} p(x_1 \mid x_2, \dots, x_k) &= p(x_1 \mid x_2) \\ p(x_k \mid x_1, \dots, x_{k-1}) &= p(x_k \mid x_{k-1}) \\ p(x_i \mid \mathbf{x} \backslash \{x_i\}) &= p(x_i \mid x_{i-1}, x_{i+1}) \end{aligned}$$

Using Problem 5.3, we have $p(x_1, \dots, x_k) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1})$. There is 1 parameter for $p(x_1)$ and 2 parameters per i for $p(x_i | x_{i-1})$. Therefore the number of parameters is $1 + 2(k-1) = \mathbf{2k} + \mathbf{1}$

3. Assume conditional independences as in the previous question. Use the chain rule of probability and the graphoid axioms to write down the likelihood for the model such that only a polynomial number of parameters (in k) are used.

Take $j \in [2, k-1]$ and suppose for $\forall i < j$ that $(X_1, \dots, X_{i-1} \perp\!\!\!\perp X_{i+1}, \dots, X_k \mid X_i)$. Then it holds that $(X_{j-1} \perp\!\!\!\perp X \backslash \{X_{j-2}, X_j\} \mid X_{j-2}, X_j) \implies (X_{j-1} \perp\!\!\!\perp X_{j+1}, \dots, X_k \mid X_1, \dots, X_{j-2}, X_j)$, $(X_1, \dots, X_{j-2} \perp\!\!\!\perp X_j, \dots, X_k \mid X_{j-1}) \implies (X_1, \dots, X_{j-2} \perp\!\!\!\perp X_{j+1}, \dots, X_k \mid X_{j-1}, X_j)$. Therefore we have that and $(X_1, \dots, X_{j-1} \perp\!\!\!\perp X_{j+1}, \dots, X_k \mid X_j)$ by the intersection axiom. Using the above and the chain rule, we can characterize the joint distribution $p(x_1, \dots, x_k)$ by $p(x_1, \dots, x_k) = \prod_{i=1}^{k} p(x_i | x_1, \dots, x_{i-1}) = \prod_{i=1}^{k} p(x_i | x_{i-1})$ Therefore $\mathcal{L}_{[D]}(\mathbf{p}) = \prod_{j=1}^{n} p_1(1)^{x_{1,j}} (1 - p_1(1))^{(1-x_{1,j})} \prod_{i=2}^{k} p_i(1|1)^{x_{i,j} x_{i-1,j}} p_i(1|0)^{x_{i,j}(1-x_{i-1,j})} (1 - p_i(1|1))^{(1-x_{i,j}) x_{i-1,j}} (1 - p_i(1|0))^{(1-x_{i,j})(1-x_{i-1,j})}$ where $p_i(1 \mid 1) = p(X_i = 1 \mid X_{i-1} = 1)$, $p_i(1 \mid 0) = p(X_i = 1 \mid X_{i-1} = 0)$ for $i \in [2, k]$ and $p_1(1) = p(X_1 = 1)$ are the parameters of the likelihood.