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: Jingguo Liang (jliang35), Partner 2: Chang Yan (cyan13)

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.

$$\begin{aligned} L_{[D]}(\omega) &= \prod_{i=1}^{n} p(y_i \mid x_i; \omega) \\ &= \prod_{i=1}^{n} p(\omega^T x_i + \epsilon_i \mid x_i; \omega) \\ &= \prod_{i=1}^{n} p(\epsilon_i \mid x_i; \omega) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - \omega^T x_i)^2}{2}\right) \\ logL_{[D]}(\omega) &= \sum_{i=1}^{n} log\frac{1}{\sqrt{2\pi}} + \sum_{i=1}^{n} -\frac{(y_i - \omega^T x_i)^2}{2} \end{aligned}$$

Maximizing $logL_{[D]}(\omega)$ is equivalent to minimizing:

$$\sum_{i=1}^{n} (y_i - \omega^T x_i)^2$$

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

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

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

$$
\begin{aligned}
L_{[D]}(\omega) &= \prod_{i=1}^{n} p(\omega^T x_i + \epsilon_i \mid x_i; \omega) \\
&= \prod_{i=1}^{n} p(\epsilon_i \mid x_i; \omega) \\
&= \prod_{i=1}^{n} \frac{1}{2} \exp\Big(- |\epsilon_i| \Big) \\
&= \prod_{i=1}^{n} \frac{1}{2} \exp\Big(- \big|y_i - \omega^T x_i\big| \Big)
\end{aligned}
$$

$$
logL_{[D]}(\omega) = \sum_{i=1}^{n} log\frac{1}{2} + \sum_{i=1}^{n}(- \big|y_i - \omega^T x_i\big|)
$$

Maximizing $logL_{[D]}(\omega)$ is equivalent to minimizing:

$$
\sum_{i=1}^{n}(\big|y_i - \omega^T x_i\big|)
$$

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

The square loss is easier to minimize: it is easier to differentiate than absolute loss, and is twice differentiable.
The absolute loss is more robust to outliers: unlike the square loss large errors does not get squared, thus magnified.

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.

$$\begin{aligned} P(cough \cap fever) &= P(cough|fever) \cdot P(fever) \\ &= 0.40 \cdot 0.20 \\ &= 0.08 \end{aligned}$$

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

$$\begin{aligned} &P(pneumonia|fever \cap no\ cough) \\ =&\frac{P(pneumonia \cap fever \cap no\ cough)}{P(fever \cap no\ cough)} \\ =&\frac{P(fever \cap no\ cough|pneumonia) \cdot P(pneumonia)}{P(no\ cough|fever) \cdot P(fever)} \\ =&\frac{P(no\ cough|fever, pneumonia) \cdot P(fever|pneumonia) \cdot P(pneumonia)}{[1 - P(cough|fever)] \cdot P(fever)} \\ =&\frac{P(no\ cough|pneumonia) \cdot P(fever|pneumonia) \cdot P(pneumonia)}{[1 - P(cough|fever)] \cdot P(fever)} \\ =&\frac{[1 - P(cough|pneumonia)] \cdot P(fever|pneumonia) \cdot P(pneumonia)}{[1 - P(cough|fever)] \cdot P(fever)} \\ =&\frac{(1 - 0.70) \cdot 0.50 \cdot 0.05}{(1 - 0.40) \cdot 0.20} \\ =&0.0625 \end{aligned}$$

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

We need to specify 5 parameters.
As $p(fever, cough, pneumonia) = p(fever, cough|pneumonia) \cdot p(pneumonia)$,
first we need 1 parameter to specify $p(pneumonia)$, and $p(no\ pneumonia)$ is just $1 - p(pneumonia)$.
Then, we know fever and cough is conditional independent given pneumonia,so we have:
$p(fever, cough|pneumonia) = p(fever|pneumonia) \cdot p(cough|pneumonia)$
$p(fever, cough|no\ pneumonia) = p(fever|no\ pneumonia) \cdot p(cough|no\ pneumonia)$
and here we need additional 4 parameters to specify them.
All the no fever/no cough ones can be calculated using 1 - one of the 4 parameters above.
So, we need a total of n = 5 parameters.

4 Conjugate Priors

1. Define what a conjugate prior is.

Conjugate priors are priors that have the same form as their updated posteriors.

2. Why are conjugate priors useful?

They reduce Bayesian updating to modifying the parameters of the prior distribution rather than computing integrals.

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^*.

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?

The size of joint probability table is $2^k - 1$.

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 $p(x_1, \dots, x_k)$?

From the assumptions we know
$p(X_1|X_{-1}) = p(X_1|X_2)$
$p(X_2|X_{-2}) = p(X_2|X_1, X_3)$
$p(X_3|X_{-3}) = p(X_3|X_2, X_4)$
...
$p(X_{k-1}|P_{-(k-1)}) = p(X_{k-1}|X_{k-2}, X_k)$
$p(X_k|X_{-k}) = p(X_k|X_{k-1})$

For $p(X_i|X_j)$, we need to specify two parameteters: $p(X_i = 1|X_j = 1)$ and $p(X_i = 1|X_j = 0)$
For $p(X_i|X_j, X_k)$, we need to specify four parameters: $p(X_i = 1|X_j = 1, X_k = 1)$, $p(X_i = 1|X_j = 1, X_k = 0)$, $p(X_i = 1|X_j = 0, X_k = 1)$, and $p(X_i = 1|X_j = 0, X_k = 0)$

Thus, a total of $4(k-2) + 4 = 4k - 4$ parameters should be specified.

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.

From $(X_1 \perp\!\!\!\perp X_3, \ldots, X_k \mid X_2)$ we have $(X_1 \perp\!\!\!\perp X_4, X_5, \ldots, X_k \mid X_2, X_3)$ (Chain rule)
And we know $(X_2 \perp\!\!\!\perp X_4, X_5, \ldots, X_k \mid X_1, X_3)$ (from (b))
Combining the two relations, we have $(X_1, X_2 \perp\!\!\!\perp X_4, X_5, \ldots, X_k \mid X_3)$ (Intersection axiom)
From there, we know $(X_1, X_2 \perp\!\!\!\perp X_5, \ldots, X_k \mid X_3, X_4)$ (Chain rule)
Combing with the assumption that $(X_3 \perp\!\!\!\perp X_1, X_5, X_6, \ldots, X_k \mid X_2, X_4)$ (from (b))
We have $(X_1, X_2, X_3 \perp\!\!\!\perp X_5, \ldots, X_k \mid X_4)$ (Intersection axiom)
And so on, we can finally have $(X_1, X_2, \ldots, X_{k-2} \perp\!\!\!\perp X_k \mid X_{k-1})$
using the conditional independences above:

$$\begin{aligned}
p(X_1, X_2, \ldots, X_k) &= p(X_1, X_3, X_4, \ldots, X_k \mid X_2)p(X_2) \\
&= p(X_1 \mid X_2)p(X_3, X_4, \ldots, X_k \mid X_2)p(X_2) \\
&= p(X_1 \mid X_2)p(X_2, X_3, \ldots, X_k) \\
&= p(X_1 \mid X_2)p(X_2 \mid X_3)p(X_4, \ldots, X_k)p(X_3) \\
&\ldots \\
&= p(X_1 \mid X_2)p(X_2 \mid X_3)\ldots p(X_{k-1} \mid X_k)p(X_k) \\
L_{[D]} = \prod_{i=1}^{n} p(x_{i1}, x_{i2}, \ldots, x_{ik}) &= \prod_{i=1}^{n} p(x_{i1} \mid x_{i2})p(x_{i2} \mid X_{i3})\ldots p(x_{i(k-1)} \mid x_{ik})p(x_k)
\end{aligned}$$

As each $p(X_i \mid X_{i+1})$ needs 2 parameters,and we have $k-1$ of them plus a $p(X_k)$ which is only 1 parameter, the likelihood only need a polynomial number of parameters $(2k-1)$ to compute.