CS181 Practice Questions: Probability and Linear Regression Basics

1. **Mean of Gaussian (Bishop, 1.8, part 1)**

   By using a change of variables, verify that the univariate Gaussian distribution satisfies

   $$\begin{aligned}\mathbb{E}[x] &= \int (2\pi\sigma^2)^{-1/2} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} x \, dx \\ &= \mu.\end{aligned}$$

   Changing variables $y = x - \mu$, gives us:

   $$\mathbb{E}[x] = \int (2\pi\sigma^2)^{-1/2} \exp\{-\frac{1}{2\sigma^2} y^2\} (y + \mu) dy.$$

2. **Mode of Gaussian (Bishop, 1.9)**

Show that the mode (i.e. the maximum) of the Gaussian distribution

$$\mathcal{N}(x|\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-(2\sigma^2)^{-1}(x-\mu)^2\}$$

is given by μ.

We differentiate with respect to x to obtain:

$$\frac{d}{dx}\mathcal{N}(x|\mu,\sigma^2) = -\mathcal{N}(x|\mu,\sigma^2)\frac{x-\mu}{\sigma^2}.$$

Setting this to zero, we obtain $x = \mu$.

3. **Gaussian MLE**

Suppose we have N iid values $x_n \sim \mathcal{N}(\mu, \sigma^2)$, where $n = 1, \ldots, N$.

(a) Write down the likelihood function.

(b) Write down the log-likelihood function.

(c) Find the maximum likelihood estimator for μ_{ML}.

(d) Find the maximum likelihood estimator for σ^2_{ML}.

(e) Show that the μ_{ML} is unbiased.

(f) Show that the σ^2_{ML} is biased.

(g) Give an unbiased estimator for the variance parameter.

(a) The likelihood function is

$$p(\boldsymbol{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2).$$

(b) The log-likelihood function is

$$\begin{aligned}
\log p(\boldsymbol{x}|\mu, \sigma^2) &= \log(\prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)) \\
&= \sum_{n=1}^{N} \log \mathcal{N}(x_n|\mu, \sigma^2) \\
&= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log(2\pi).
\end{aligned}$$

(c) To find the MLE, we take the derivative with respect to μ and set it equal to zero. Solving for σ^2, we get that the MLE of the mean estimator is the sample mean:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n.$$

(d) To find the MLE, we take the derivative with respect to σ^2 and set it equal to zero. Solving for σ^2, we get that the MLE of the variance estimator is the sample variance:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu})^2.$$

(e) The bias for the mean estimator is:

$$\begin{aligned}
\text{bias}(\mu_{\text{ML}}) &= \mathbb{E}[\mu_{\text{ML}}] - \mu \\
&= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[x_n] - \mu \\
&= \mu - \mu = 0.
\end{aligned}$$

(f) The bias for the variance estimator is:

$$\begin{aligned}
\text{bias}(\sigma^2_{\text{ML}}) &= \mathbb{E}[\sigma^2_{\text{ML}}] - \sigma^2 \\
&= \left(\frac{N-1}{N}\right) \sigma^2 - \sigma^2 \\
&= \frac{N-1}{N} - 1.
\end{aligned}$$

(g) To get an unbiased estimator for the variance, we multiply by the bias:

$$\frac{N}{N-1} \sigma^2_{ML}.$$

4. **MLE Estimate of the Bias Term (Bishop (3.19))**

Let $\boldsymbol{\Phi}$ be our $N \times J$ design matrix, $\boldsymbol{t}$ our vector of N target values, $\boldsymbol{w}$ our vector of J parameters, and w_0 our bias parameter. As Bishop notes in (3.18), the sum-of-squares error function of $\boldsymbol{w}$ and w_0 can be written as follows

$$E(\boldsymbol{w}, w_0) = \frac{1}{2} \sum_{n=1}^{N} \left(t_n - w_0 - \sum_{j=1}^{J-1} w_j \cdot \phi_j(x_n) \right)^2 .$$

Show that the value of w_0 that minimizes E is

$$\begin{aligned} w_{0_{MLE}} &= \frac{1}{N} \sum_{n=1}^{N} t_n - \sum_{j=1}^{J-1} w_j \cdot \left(\frac{1}{N} \sum_{n=1}^{N} \phi_j(x_n) \right) \\ &= \bar{t} - \sum_{j=1}^{J-1} w_j \cdot \overline{\phi_j(x)} \qquad \text{[compare Bishop (3.19)]} \end{aligned}$$

We have that $\frac{\partial E}{\partial w_0} = -\sum_{n=1}^{N} (t_n - w_0 - \sum_{j=1}^{J-1} w_j \cdot \phi_j(x_n))$.

Thus, we set $\sum_{n=1}^{N} t_n - N w_0 - \sum_{n=1}^{N} \sum_{j=1}^{J-1} w_j \cdot \phi_j(x_n) = 0$, and solving for w_0 gives the result.

5. **Simple Linear Regression (Bishop, 1.1)**

Consider the sum-of-squares error function given by:

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2,$$

in which the function $y(x, \boldsymbol{w})$ is given by the polynomial

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j.$$

Show that the coefficients $\boldsymbol{w} = \{w_i\}$ that minimize this error function are given by the solution to the following set of linear equations

$$\sum_{j=0}^{M} A_{ij} w_j = T_i$$

where

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j},$$
$$T_i = \sum_{n=1}^{N} (x_n)^i t_n.$$

Here a suffix i or j denotes the index of a component, where as $(x)^i$ denotes x reaised to the power of i.

Substituting the second equation into the first equation and then differentiating with respect to w_i, we obtain

$$\sum_{n=1}^{N} \left(\sum_{j=0}^{M} w_j x_n^j - t_n \right) x_n^i = 0.$$

6. **Multivariate Regression (Adapted from Stanford CS229)**

Suppose we have $\boldsymbol{\Phi} \in \mathbb{R}^{N \times J}$ as our design matrix, but that instead of predicting scalar values t_n, we'd like to use least squares regression to predict vector-valued targets $\boldsymbol{t}_n \in \mathbb{R}^M$ for each row $\phi(\boldsymbol{x}_n)$ in $\boldsymbol{\Phi}$. To do this, we can introduce a parameter matrix $\boldsymbol{W} \in \mathbb{R}^{N \times M}$ and attempt to minimize the following sum-of-squares error function:

$$
\begin{aligned}
E(\boldsymbol{W}) &= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}\left((\boldsymbol{W}^\top \phi(\boldsymbol{x}_n))_m - t_{nm}\right)^2 \\
&= \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{W}^\top \phi(\boldsymbol{x}_n) - \boldsymbol{t}_n)^\top(\boldsymbol{W}^\top \phi(\boldsymbol{x}_n) - \boldsymbol{t}_n) \\
&= \frac{1}{2}\mathrm{tr}((\boldsymbol{\Phi}\boldsymbol{W} - \boldsymbol{T})^\top(\boldsymbol{\Phi}\boldsymbol{W} - \boldsymbol{T})),
\end{aligned}
$$

where $\boldsymbol{T} = \begin{bmatrix} — & \boldsymbol{t}_1^\top & — \\ & \vdots & \\ — & \boldsymbol{t}_N^\top & — \end{bmatrix}$.

(a) Rewrite $E(\boldsymbol{W})$ in terms of the traces of 3 matrices.

(b) Derive the gradient of your answer to part (a) with respect to $\boldsymbol{W}$.

(c) Derive $\boldsymbol{W}_{MLE}$ by setting the gradient of your answer to part (b) to 0.

Hint: in answering the above questions, it may be useful to keep in mind the following facts:

1. $\mathrm{tr}(\boldsymbol{A} + \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B})$
2. $\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}^\top)$
3. $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A})$
4. $\frac{\partial}{\partial \boldsymbol{A}}\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \boldsymbol{B}^\top$
5. $\frac{\partial}{\partial \boldsymbol{A}}\mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{B}) = \boldsymbol{B}$

(a)

$$
\begin{aligned}
\frac{1}{2}\mathrm{tr}((\boldsymbol{\Phi}\boldsymbol{W} - \boldsymbol{T})^\top(\boldsymbol{\Phi}\boldsymbol{W} - \boldsymbol{T})) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{W}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{W} - \boldsymbol{W}^\top\boldsymbol{\Phi}^\top\boldsymbol{T} - \boldsymbol{T}^\top\boldsymbol{\Phi}\boldsymbol{W} + \boldsymbol{T}^\top\boldsymbol{T}) \\
&= \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{W}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{W}) - \mathrm{tr}(\boldsymbol{W}^\top\boldsymbol{\Phi}^\top\boldsymbol{T}) - \mathrm{tr}(\boldsymbol{T}^\top\boldsymbol{\Phi}\boldsymbol{W}) + \mathrm{tr}(\boldsymbol{T}^\top\boldsymbol{T})\right] \\
&= \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{W}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{W}) - 2\mathrm{tr}(\boldsymbol{T}^\top\boldsymbol{\Phi}\boldsymbol{W}) + \mathrm{tr}(\boldsymbol{T}^\top\boldsymbol{T})\right],
\end{aligned}
$$

where in the last line we used the fact that $\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}^\top)$.

(b) We have $\nabla_W E = \frac{1}{2}\left[\mathbf{\Phi}^\top \mathbf{\Phi} \boldsymbol{W} + \mathbf{\Phi}^\top \mathbf{\Phi} \boldsymbol{W} - 2\mathbf{\Phi}^\top \boldsymbol{T}\right] = \mathbf{\Phi}^\top \mathbf{\Phi} \boldsymbol{W} - \mathbf{\Phi}^\top \boldsymbol{T}$, where the first equality is obtained by using the product rule on the first trace, and the 4th identity on the second trace.

(c) Setting our answer to 0, we get $\boldsymbol{W}_{MLE} = (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \boldsymbol{T}$, which is almost exactly the same as in the scalar prediction case.