Math 789 Assignment 3

Brendan Drachler

February 13, 2020

Bishop 2.1

Beginning from Eq. 2.1,

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{1}$$

We want to use the fact that:

$$\int_{-\infty}^{\infty} e^{\frac{\lambda}{2}x^2} dx = (\frac{2\pi}{\lambda})^{\frac{1}{2}} \tag{2}$$

To verify Eq. 2.2 and 2.3 as well as to prove that $\int p(x)dx = 1$. I'll start by proving the latter.

By defining $\lambda = \frac{1}{\sigma^2}$ and integrating Eq. 1, we can rewrite it as,

$$\int p(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2}\lambda} = (\frac{2\pi}{\lambda})^{\frac{1}{2}} (\frac{1}{2\pi})^{\frac{1}{2}} (\frac{1}{\sigma}) \tag{3}$$

$$\downarrow \tag{4}$$

$$(\frac{1}{\sigma^2})^{\frac{1}{2}} (\frac{1}{\sigma}) = 1 \tag{5}$$

Now, to prove Eq. 2.2 in Bishop, we have to solve.

$$\int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx \tag{6}$$

$$\downarrow \tag{7}$$

$$\frac{e^{-(x-\mu)^2/(2\sigma^2)} + (\frac{\pi}{2})^{\frac{1}{2}} \mu\sigma erf(\frac{x-\mu}{\sqrt{2}\sigma})}{\sqrt{2\pi}\sigma}\Big|_{-\infty}^{\infty} = \mu \tag{8}$$

This validates Eq. 2.2. Now for Eq. 2.3:

$$\int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx \tag{9}$$

$$\downarrow \tag{10}$$

$$\frac{1}{2}\sigma(\sigma erf(\frac{x-\mu}{\sqrt{2}\sigma}) + \sqrt{\frac{2}{\pi}}(\mu - x)e^{-(x-\mu)^2/(2\sigma^2)})\Big|_{-\infty}^{\infty} = \frac{1}{2}\sigma(\sigma+\sigma) = \sigma^2 \tag{11}$$

This validates Eq. 2.3 as well!

Bishop 2.3

Starting from Eq. 2.1:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{12}$$

and utilizing

$$E = -\ln L(\theta) = -\sum_{n=1}^{N} \ln P(x^n|\theta) \tag{13}$$

Plugging in $P(x)$, we can simplify E to:

$$E = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{j-1}^{N}(x_j - \mu)^2 \tag{14}$$

We can now differentiate with respect to the two variables, μ and σ^2.

$$\frac{\partial E}{\partial \mu} = \frac{1}{2\sigma^2}\sum_{j-1}^{N} 2(x_j - \mu) = 0 \tag{15}$$

$$\sum_{j-1}^{N} x_j - \sum_{j-1} Nn\mu = 0 \tag{16}$$

$$\sum_{j-1}^{N} x_j - N\mu = 0 \tag{17}$$

$$\frac{1}{N}\sum_{j-1}^{N} x_j = \mu \tag{18}$$

This reproduces Bishop Eq. 2.21. Now we can differentiate with respect to σ^2.

$$\frac{\partial E}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j-1}^{N}(x_j - \mu)^2 = 0 \tag{19}$$

Multiplying by σ^2:

$$\frac{N\sigma^2}{2} - \frac{1}{2}\sum_{j-1}^{n}(x_j - \mu)^2 = 0 \tag{20}$$

$$\sigma^2 = \frac{1}{N}\sum_{j-1}^{n}(x_j - \mu)^2 \tag{21}$$

This validates Bishop Eq. 2.22.

Bishop 2.5

As usual, we will begin by assuming a distribution given by Bishop Eq. 2.1:

$$p(x^n|\mu) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2} \tag{22}$$

and a prior distribution given by:

$$p_0(\mu) = \frac{1}{\sigma_o\sqrt{2\pi}}e^{-(\mu-\mu_o)^2/2\sigma_o^2} \tag{23}$$

We know that the product of these two will result in an exponential term and a constant term. My goal is to only analyze what will happen inside the exponential term because if are able to get it in the form $exp(-\frac{1}{2}\frac{\mu-\mu_N}{\sigma_N^2})$, we can deduce what μ_N and σ_N^2 are. Multiplying the two distributions above, calling the catch all constant out front, A, and focusing on the $exp()$ term will result in:

$$p_N(\mu|\chi) = A\ exp(-\frac{(\mu-\mu_o)^2}{2\sigma_o^2} + \sum_{n=1}^{N} -\frac{(x^n-\mu)^2}{2\sigma^2}) \tag{24}$$

Any term not containing a μ is in essence a multiplicative term that we will wrap up into A. With this in mind, we will foil and hide all unnecessary terms.

$$p_N(\mu|\chi) = A\ exp(-\frac{1}{2}\frac{(\sigma^2 + N\sigma_o^2)\mu^2 - 2(\sigma^2\mu_o + N\overline{x}\sigma_o^2)\mu}{\sigma^2\sigma_o^2}) \tag{25}$$

We want to get rid of the coefficient in front of the μ^2 term so we can complete the square easily.

$$p_N(\mu|\chi) = A\ exp(-\frac{1}{2}\frac{\mu^2 - 2\frac{(\sigma^2\mu_o + N\overline{x}\sigma_o^2)}{(\sigma^2 + N\sigma_o^2)}\mu}{\frac{\sigma^2\sigma_o^2}{\sigma^2 + N\sigma_o^2}}) \tag{26}$$

Completing the square in the numerator leads to:

$$p_N(\mu|\chi) = A\ exp(-\frac{1}{2}\frac{(\mu - \frac{N\overline{x}\sigma_o^2 + \sigma^2\mu_o}{\sigma^2 + N\sigma_o^2})^2}{\frac{\sigma_o^2\sigma^2}{\sigma^2 + N\sigma_o^2}}) \tag{27}$$

We finally have this in the form we need! The term subtracted from the μ is μ_N and the term in the denominator is σ_N^2.

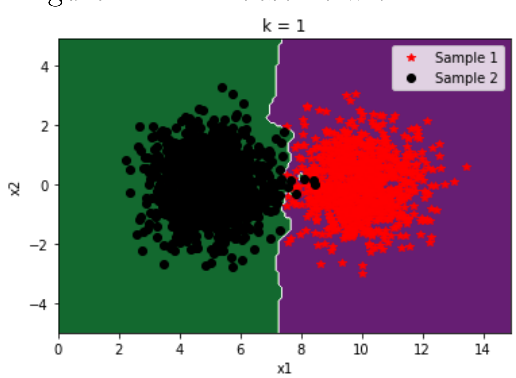
$$\mu_N = \frac{N\overline{x}\sigma_o^2 + \sigma^2\mu_o}{\sigma^2 + N\sigma_o^2}\ ,\ \sigma_N^2 = \frac{\sigma_o^2\sigma^2}{\sigma^2 + N\sigma_o^2} \tag{28}$$

Bishop 2.9

I've generated two multivariate samples with the covariance matrix equal to the identity and the mean of each sample being, $\mu_1 = 10$ and $\mu_2 = 5$.

My K-nearest neighbor predictor for different values of K are below. The $k = 1$ case seems to be overfitting severely. It is making an effort to separate every point. It likely could not be generalized.

Figure 1: KNN best fit with k = 1.



The $k = 5$ case does a good job of separating the two populations with good generality.

The $k = 20$ case seems to be averaging the populations which is arguably not the best fit.

Figure 2: KNN best fit with k = 5.

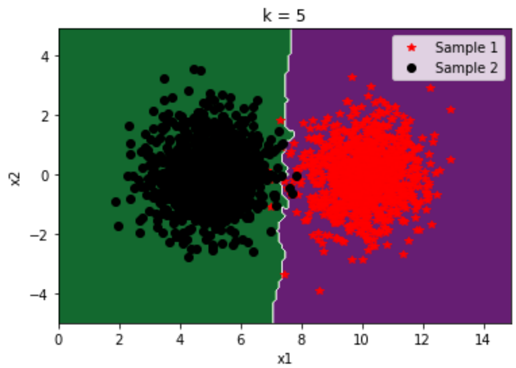
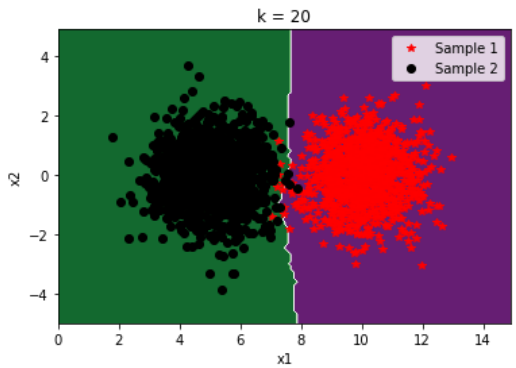


Figure 3: KNN best fit with k = 20.



Bishop 2.10

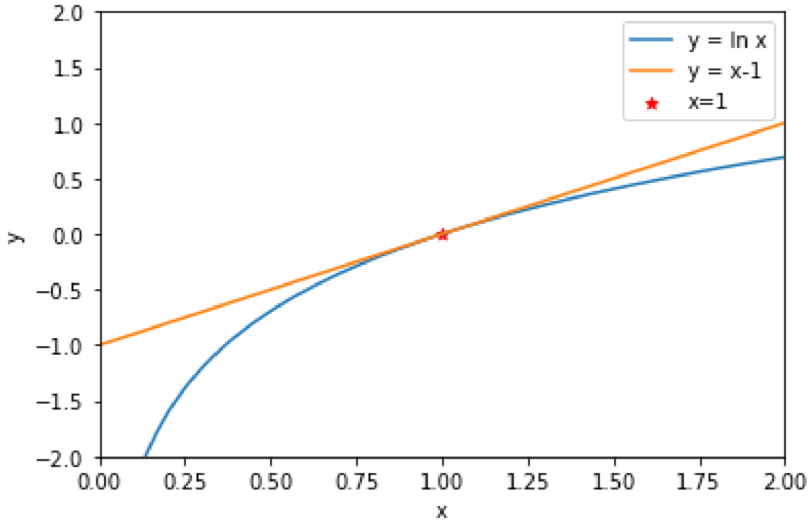
Now, we will maximize $\ln x - \ x - 1$.The inequality says this should always be negative expect at $x = 1$. Therefore, we expect the maximum to be at $x = 1$.

$$\frac{d}{dx}(\ln x - \ x - 1) = 0 \tag{29}$$

$$\frac{1}{x} - 1 = 0 \tag{30}$$

$$x = 1 \tag{31}$$

Figure 4: Verifies the inequality $\ln x \leq \; x - 1$



Now, we will analyze the Kullback-Leibler distance is given by:

$$L = -\int p(x) \ln \frac{\tilde{p}(x)}{p(x)} dx \tag{32}$$

We will draw a parallel with the first part of this exercise by saying $\frac{\tilde{p}(x)}{p(x)} = x$.

$$\int p(x) \ln \frac{\tilde{p}(x)}{p(x)} dx \leq \int p(x) (\frac{\tilde{p}(x)}{p(x)} - 1) dx \tag{33}$$

Differentiating both sides leads to:

$$\ln \frac{\tilde{p}(x)}{p(x)} \leq (\frac{\tilde{p}(x)}{p(x)} - 1) \tag{34}$$

$$\ln \frac{\tilde{p}(x)}{p(x)} - (\frac{\tilde{p}(x)}{p(x)} - 1) \leq 0 \tag{35}$$

Find the maximum, which is identical to what was done above:

$$\tilde{p}(x) = p(x) \tag{36}$$

Therefore, $L \geq 0$ with equality if $\tilde{p}(x) = p(x)$.

Bishop 2.11

Applying the Lagrangian to this case with the constraint, $\sum_i q_i = 1$.

$$-\sum_i p_i \ln(\frac{q_i}{p_i}) + \lambda(\sum_i q_i - 1) = 0 \tag{37}$$

$$-\sum_i p_i \ln(q_i) + \sum_i p_i \ln(p_i) + \lambda \sum_i q_i - \lambda = 0(p_i + q_i + \lambda) \tag{38}$$

Now, I will use a undetermined coefficients type method to solve this.

$$\lambda \sum_i q_i - \lambda = 0 \tag{39}$$

$$\sum_i q_i = 1 \tag{40}$$

This proves our constraint. Now to prove the second part,

$$-\ln(q_i) + \ln(p_i) = 0 \tag{41}$$

$$q_i = p_i \tag{42}$$

We've shown in the previous problem that if two discrete distributions are equal, the Kullback-Leibler distance is 0. And in this problem we proved that $p_i = q_i$. Therefore, the KL distance between them is 0.