Assignment 1

Nicholai L'Esperance

06/05/20

1 Question 1

1.1 By using a change of variables, verify that the univariate Gaussian distribution given by$N\left(x|\mu,\sigma^2\right)$ satisfies $E\left(x\right)=\mu$

The formula for expected value is $E\left(x\right)=\int_{-\infty}^{\infty}xf\left(x\right)$. Thus:

$$N\left(x|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E\left(N\left(x|\mu,\sigma^2\right)\right)=\int_{-\infty}^{\infty}x\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

We can solve this using substitution. For this, we define a new variable, v:

$$v=\frac{x-\mu}{\sqrt{2}\sigma}$$

Differentiating, we have

$$\frac{dv}{dx}=\frac{d}{dx}\left(\frac{x}{\sqrt{2}\sigma}-\frac{\mu}{\sqrt{2}\sigma}\right)=\frac{1}{\sqrt{2}\sigma}$$

For our substituation, we must solve for x in terms of v.

$$x=\sqrt{2}\sigma v+\mu$$

Next, we can substitute in v, and we have:

$$E\left(N\left(x|\mu,\sigma^2\right)\right)=\int_{-\infty}^{\infty}x\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}=\int_{-\infty}^{\infty}\left(\sqrt{2}\sigma v+\mu\right)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-v^2}dx$$

Finally, we can substitute in dv for dx, noting $dx=2\sigma dv$.

$$E\left(N\left(x|\mu,\sigma^2\right)\right)=\int_{-\infty}^{\infty}\left(\sqrt{2}\sigma v+\mu\right)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-v^2}\left(2\sigma\right)dv$$

Simplifying:

$$\int_{-\infty}^{\infty}\left(\sqrt{2}\sigma v+\mu\right)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-v^2}\left(2\sigma\right)dv=\int_{-\infty}^{\infty}\frac{2\sigma v}{\sqrt{\pi}}e^{-v^2}+\frac{2\mu}{\sqrt{2\pi}}e^{-v^2}dv=\frac{2}{\sqrt{\pi}}\left(\sigma\int_{-\infty}^{\infty}ve^{-v^2}dv+\frac{\mu}{\sqrt{2}}\int_{-\infty}^{\infty}e^{-v^2}dv\right)$$

We can evaluate each of the integrals individually. First

$$\sigma\int_{-\infty}^{\infty}ve^{-v^2}=\sigma\left(-\frac{1}{2}e^{-x^2}\Big|_{-\infty}^{\infty}\right)=\sigma\left(0-0\right)=0$$

The second integral is trickier, but has a known solution. The following integral is known as the "Gaussian Integral":

$$\frac{\mu}{\sqrt{2}}\int_{-\infty}^{\infty}e^{-v^2}dv=\frac{\mu}{\sqrt{2}}\sqrt{\pi}$$

We can now arrive at our final solution:

$$E\left(N\left(x|\mu,\sigma^2\right)\right)=\frac{2}{\sqrt{\pi}}\left(\sigma\int_{-\infty}^{\infty}ve^{-v^2}dv+\frac{\mu}{\sqrt{2}}\int_{-\infty}^{\infty}e^{-v^2}dv\right)=\frac{2}{\sqrt{2\pi}}\left(0+\frac{\mu}{\sqrt{2}}\sqrt{\pi}\right)=\frac{2\mu\sqrt{\pi}}{2\sqrt{\pi}}=\mu$$

1.2 Next, by differentiating both sides of normalization condition $\int_{-\infty}^{\infty} N\left(x|\mu,\sigma^2\right) = 1$ with respect to σ^2, verify that the Gaussian satisfies $E\left(x^2\right) = \mu^2 + \sigma^2$.

Our expected value formula is defined as:

$$E\left(x^2\right) = \int_{-\infty}^{\infty} x^2 N\left(x|\mu,\sigma^2\right)$$

First, we take our function N, and find the derivative w.r.t. σ. This could also be done w.r.t. σ^2 via a change of variables, but I find it simpler to keep in terms of σ. Because we have an exponential, it is easiest to take the natural log of both sides, first.

$$N\left(x|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$ln\left(N\left(x|\mu,\sigma^2\right)\right) = ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

$$\frac{\delta}{\delta\sigma}\left(ln\left(N\left(x|\mu,\sigma^2\right)\right)\right) = \frac{\delta}{\delta\sigma}\left(ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + ln\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)\right)$$

We solve the derivative of the natural log using the chain rule.

$$\frac{1}{N\left(x|\mu,\sigma^2\right)}\frac{\delta N\left(x|\mu,\sigma^2\right)}{\delta\sigma} = \frac{\delta}{\delta\sigma}\left(-ln\left(\sqrt{2\pi\sigma^2}\right) - \frac{(x-\mu)^2}{2\sigma^2}\right)$$

We again use the chain rule, to solve the derivative of the natural log on the right hand side.

$$\frac{1}{N\left(x|\mu,\sigma^2\right)}\frac{\delta N\left(x|\mu,\sigma^2\right)}{\delta\sigma} = -\frac{\sqrt{2\pi}}{\sqrt{2\pi\sigma^2}} + \frac{2\left(x-\mu\right)^2}{2\sigma^3} = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}$$

Finally, we have

$$\frac{\delta N\left(x|\mu,\sigma^2\right)}{\delta\sigma} = \left(\frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}\right)N\left(x|\mu,\sigma^2\right)$$

Taking a look at our normalization formula, we note the following:

$$\int_{-\infty}^{\infty} N\left(x|\mu,\sigma^2\right)dx = 1$$

$$\int_{-\infty}^{\infty} \frac{\delta}{\delta\theta}\left(N\left(x|\mu,\sigma^2\right)\right)dx = \frac{\delta}{\delta\theta}\left(1\right) = 0$$

We can plug in our equations from above

$$\int_{-\infty}^{\infty} \frac{\delta}{\delta\theta}\left(N\left(x|\mu,\sigma^2\right)\right)dx = \int_{-\infty}^{\infty}\left(\frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}\right)N\left(x|\mu,\sigma^2\right)dx = 0$$

$$\int_{-\infty}^{\infty} \frac{x^2}{\sigma^3}N\left(x|\mu,\sigma^2\right) - \frac{2x\mu}{\sigma^3}N\left(x|\mu,\sigma^2\right) + \frac{\mu^2}{\sigma^3}N\left(x|\mu,\sigma^2\right) - \frac{1}{\sigma}N\left(x|\mu,\sigma^2\right)dx = 0$$

We can now recognize the expected value formula here

$$\frac{1}{\sigma^3}E\left(x^2\right) - \frac{2\mu}{\sigma^3}E\left(x\right) + \int_{-\infty}^{\infty}\frac{\mu^2}{\sigma^3}N\left(x|\mu,\sigma^2\right)dx - \int_{-\infty}^{\infty}\frac{1}{\sigma}N\left(x|\mu,\sigma^2\right)dx = 0$$

The remaining two integrals are solved trivially using the normalization condition above. Also, the expected value of x is defined as μ.

$$\frac{1}{\sigma^3}E\left(x^2\right) - \frac{2\mu}{\sigma^3}\mu + \frac{\mu^2}{\sigma^3} - \frac{1}{\sigma} = 0$$

Now, we solve for $E\left(x^2\right)$

$$\frac{1}{\sigma^3}E\left(x^2\right) - \frac{\mu^2}{\sigma^3} = \frac{1}{\sigma}$$

$$\frac{1}{\sigma^3}E\left(x^2\right) = \frac{1}{\sigma} + \frac{\mu^2}{\sigma^3}$$

$$E\left(x^2\right) = \frac{\sigma^3}{\sigma} + \frac{\sigma^3\mu^2}{\sigma^3} = \sigma^2 + \mu^2$$

Thus, we have proven that $E\left(x^2\right) = \sigma^2 + \mu^2$.

2 Question 2

2.1 Use $E(x) = \mu$ to prove $E(xx^T) = \mu\mu^T + \Sigma$

We know

$$E(\mathbf{x}) = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \dots \\ E(x_n) \end{bmatrix} \equiv \mu,\ E\left(x^T\right) = \begin{bmatrix} E(x_1) & E(x_2) & \dots & E(x_n) \end{bmatrix} \equiv \mu^T$$

We have two matrices to evaluate. First

$$E\left(xx^T\right) = E\left(\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}\right) = E\left(\begin{bmatrix} x_1^2 & x_1x_2 & \dots & x_1x_n \\ x_1x_2 & x_2^2 & & \\ \dots & & \dots & \\ x_1x_n & & & x_n^2 \end{bmatrix}\right)$$

$$= \begin{bmatrix} E\left(x_1^2\right) & E(x_1x_2) & \dots & E(x_1x_n) \\ E(x_1x_2) & E\left(x_2^2\right) & & \\ \dots & & \dots & \\ E(x_1x_n) & & & E\left(x_n^2\right) \end{bmatrix}$$

Next

$$\mu\mu^T = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \dots \\ E(x_n) \end{bmatrix} \begin{bmatrix} E(x_1) & E(x_2) & \dots & E(x_n) \end{bmatrix} = \begin{bmatrix} E(x_1)^2 & E(x_1)E(x_2) & \dots & E(x_1)E(x_n) \\ E(x_1)E(x_2) & E(x_2)^2 & & \dots \\ \dots & & \dots & E(x_{n-1})E(x_n) \\ E(x_1)E(x_n) & \dots & E(x_{n-1})E(x_n) & E(x_n)^2 \end{bmatrix}$$

The last piece of the puzzle is the covariance matrix. This is defined as

$$\Sigma = E\left(xx^T\right)\text{-}E(x)\,E\left(x^T\right)$$

This is trivially calculated from the results above

$$\Sigma = \begin{bmatrix} E\left(x_1^2\right) - E(x_1)^2 & E(x_1x_2) - E(x_1)E(x_2) & \dots & E(x_1x_n) - E(x_1)E(x_n) \\ E(x_1x_2) - E(x_1)E(x_2) & E\left(x_2^2\right) - E(x_2)^2 & & \\ \dots & & \dots & \\ E(x_1x_n) - E(x_1)E(x_n) & & & E\left(x_n^2\right) - E(x_n)^2 \end{bmatrix}$$

Finally, we arrive at the final solution.

$$\mu\mu^T + \Sigma = \begin{bmatrix} E(x_1)^2 + E\left(x_1^2\right) - E(x_1)^2 & E(x_1)E(x_2) + E(x_1x_2) - E(x_1)E(x_2) & \dots & E(x_1)E(x_n) + E(x_1x_n) - E(x_1)E(x_n) \\ E(x_1)E(x_2) + E(x_1x_2) - E(x_1)E(x_2) & E(x_2)^2 + E\left(x_2^2\right) - E(x_2)^2 & & \\ \dots & & \dots & \\ E(x_1)E(x_n) + E(x_1x_n) - E(x_1)E(x_n) & & & E(x_n)^2 + E\left(x_n^2\right) - E(x_n)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E\left(x_1^2\right) & E(x_1x_2) & \dots & E(x_1x_n) \\ E(x_1x_2) & E\left(x_2^2\right) & & \\ \dots & & \dots & \\ E(x_1x_n) & & & E\left(x_n^2\right) \end{bmatrix} = E\left(xx^T\right)$$

2.2 Now, using the results two definitions, show that $E[x_nx_m] = \mu\mu^T + I_{nm}\Sigma$

We prove this for the general case, with x being k values long. We can plug our calculated values in directly to prove this equality

$$\mu\mu^T + I_{nm}\Sigma = \begin{bmatrix} E(x_1)^2 & E(x_1)E(x_2) & \dots & E(x_1)E(x_k) \\ E(x_1)E(x_2) & E(x_2)^2 & & \dots \\ \dots & & \dots & E(x_{k-1})E(x_k) \\ E(x_1)E(x_k) & \dots & E(x_{k-1})E(x_k) & E(x_k)^2 \end{bmatrix} +$$

$$\begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & & \\ ... & & ... & \\ 0 & & & 1 \end{bmatrix} \begin{bmatrix} E\left(x_1^2\right) - E\left(x_1\right)^2 & E\left(x_1x_2\right) - E\left(x_1\right)E\left(x_2\right) & ... & E\left(x_1x_k\right) - E\left(x_1\right)E\left(x_k\right) \\ E\left(x_1x_2\right) - E\left(x_1\right)E\left(x_2\right) & E\left(x_2^2\right) - E\left(x_2\right)^2 & & \\ ... & & ... & \\ E\left(x_1x_k\right) - E\left(x_1\right)E\left(x_k\right) & & & E\left(x_k^2\right) - E\left(x_k\right)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E\left(x_1\right)^2 & E\left(x_1\right)E\left(x_2\right) & ... & E\left(x_1\right)E\left(x_n\right) \\ E\left(x_1\right)E\left(x_2\right) & E\left(x_2\right)^2 & & ... \\ ... & & ... & E\left(x_{k-1}\right)E\left(x_k\right) \\ E\left(x_1\right)E\left(x_{nk}\right) & ... & E\left(x_{k-1}\right)E\left(x_k\right) & E\left(x_k\right)^2 \end{bmatrix} +$$

$$\begin{bmatrix} E\left(x_1^2\right) - E\left(x_1\right)^2 & 0 & ... & 0 \\ 0 & E\left(x_2^2\right) - E\left(x_2\right)^2 & & ... \\ ... & & ... & 0 \\ 0 & ... & 0 & E\left(x_k^2\right) - E\left(x_k\right)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E\left(x_1\right)^2 + E\left(x_1^2\right) - E\left(x_1\right)^2 & E\left(x_1\right)E\left(x_2\right) & ... & E\left(x_1\right)E\left(x_n\right) \\ E\left(x_1\right)E\left(x_2\right) & E\left(x_2\right)^2 + E\left(x_2^2\right) - E\left(x_2\right)^2 & & ... \\ ... & & ... & E\left(x_{k-1}\right)E\left(x_k\right) \\ E\left(x_1\right)E\left(x_k\right) & ... & E\left(x_{k-1}\right)E\left(x_k\right) & E\left(x_k\right)^2 + E\left(x_k^2\right) - E\left(x_k\right)^2 \end{bmatrix}$$

And finally, we see

$$= \begin{bmatrix} E\left(x_1^2\right) & E\left(x_1\right)E\left(x_2\right) & ... & E\left(x_1\right)E\left(x_n\right) \\ E\left(x_1\right)E\left(x_2\right) & E\left(x_2^2\right) & & ... \\ ... & & ... & E\left(x_{k-1}\right)E\left(x_k\right) \\ E\left(x_1\right)E\left(x_k\right) & ... & E\left(x_{k-1}\right)E\left(x_k\right) & E\left(x_k^2\right) \end{bmatrix} = E\left(xx^T\right)$$

And thus

$$E[x_n x_m] = \mu\mu^T + I_{nm}\Sigma$$

3 Question 3

Show that minimizing L_D averaged over the noise distribution is equivalent to minimizing the sum of square error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameters w_0 is omitted from the regularizer.

$$L_{old}(w) = \frac{1}{2}\sum_n \left(f(x) - y_n\right)^2$$

$$L_{new}(w) = \frac{1}{2}\sum_n \left(\left(w_0 + \Sigma_i w_i\left(x_i + \epsilon_i\right)\right) - y_n\right)^2$$

rewriting as f(x), for simplicity

$$L_{new}(w) = \frac{1}{2}\sum_n \left(\left(f(x) + \Sigma_i w_i \epsilon_i\right) - y_n\right)^2$$

expanding the quadratic

$$L_{new}(w) = \frac{1}{2}\sum_n \left(f(x)^2 + 2f(x)\sum_i w_i\epsilon_i + \left(\sum_i w_i\epsilon_i\right)^2 - 2f(x)y_n - 2y_n\left(\sum_i w_i\epsilon_i\right) + y_n^2\right)$$

we can re-arrange the terms to pull out our old loss function

$$L_{new}(w) = \frac{1}{2}\sum_n \left(\left(f(x)^2 - 2f(x)y_n + y_n^2\right) + 2f(x)\sum_i w_i\epsilon_i + \left(\sum_i w_i\epsilon_i\right)^2 - 2y_n\left(\sum_i w_i\epsilon_i\right)\right)$$

rewriting with the old Loss function, we still have a number of terms that include f(x). All of our noise terms also include w.

$$L_{new}(w) = L_{old}(w) + \frac{1}{2}\sum_n \left(2f(x)\sum_i w_i\epsilon_i + \left(\sum_i w_i\epsilon_i\right)^2 - 2y_n\left(\sum_i w_i\epsilon_i\right)\right)$$

Now, we can can take the expected value of both side of the equation.

$$E(L_{new}\left(w\right)) = E\left(L_{old}\left(w\right)\right) + E\left(\frac{1}{2}\sum_n\left(2f\left(x\right)\sum_i w_i\epsilon_i + \left(\sum_i w_i\epsilon_i\right)^2 - 2y_n\left(\sum_i w_i\epsilon_i\right)\right)\right)$$

$$= E\left(L_{old}\left(w\right)\right) + \frac{1}{2}\sum_n\left(2f\left(x\right)E\left(\sum_i w_i\epsilon_i\right) + E\left(\left(\sum_i w_i\epsilon_i\right)^2\right) - 2y_n E\left(\sum_i w_i\epsilon_i\right)\right)$$

We know that $E\left(\epsilon_i\right) = 0$, which eliminates several terms. However, we must also calculate $E\left(\left(\sum_i w_i\epsilon_i\right)^2\right)$. From algebra, we know this can be broken out into two summations.

$$\left(\sum_i w_i\epsilon_i\right)^2 = \sum_i \left(w_i\epsilon_i\right)^2 + \sum_j\sum_{i\neq j} w_i\epsilon_i w_j\epsilon_j$$

Now, we can take the expected value of both sides. Note, the expected value does not depend on w, so we can just take the expected value of the epsilon noise terms.

$$E\left(\left(\sum_i w_i\epsilon_i\right)^2\right) = E\left(\sum_i \left(w_i\epsilon_i\right)^2\right) + E\left(2\sum_j\sum_{i\neq j} w_i\epsilon_i w_j\epsilon_j\right) = \sum_i w_i^2 E\left(\epsilon_i^2\right) + \sum_j\sum_{i\neq j} w_i w_j E\left(\epsilon_i\epsilon_j\right)$$

From our definitions of gaussian noise:

$$E\left(\left(\sum_i w_i\epsilon_i\right)^2\right) = \sum_i w_i^2\sigma^2 + 2\sum_j\sum_{i\neq j} w_i w_j\sigma^2 = \sigma^2\left(\sum_i w_i^2 + \sum_j\sum_{i\neq j} w_i w_j\right) = \sigma^2\left(\sum_j\sum_i w_i w_j\right)$$

Now, we can plug these results back into our expected value formula.

$$E(L_{new}\left(w\right)) = E\left(L_{old}\left(w\right)\right) + \frac{1}{2}\sum_n\left(\cancelto{0}{2f\left(x\right)E\left(\sum_i w_i\epsilon_i\right)} + E\left(\left(\sum_i w_i\epsilon_i\right)^2\right) \cancelto{0}{- 2y_n E\left(\sum_i w_i\epsilon_i\right)}\right)$$

$$= E\left(L_{old}\left(w\right)\right) + \frac{1}{2}\sum_n\left(\sigma^2\left(\sum_j\sum_i w_i w_j\right)\right)$$

Now, there are no terms in our summatation dependent on n, we we can simplify. Additionally, we know that the expected value of our L_{old} function is, by definition, the average (simply divide by N).

$$E(L_{new}\left(w\right)) = \frac{L_{old}\left(w\right)}{N} + \frac{N\sigma^2}{2}\left(\sum_j\sum_i w_i w_j\right)$$

Now, we can see that the expected value of our new Loss function is equal to the average (or expecte value of) the old loss function, added to a new 'regularization' term (which does not include w_0). <u>*Note:*</u> Professor, you mentioned there should not be an 'N' factor in the second term. However, according to my maths above, I do not see a way to get rid of this N factor.