Assignment 1

Homework assignments will be done individually: each student must hand in their own answers. Use of partial or entire solutions obtained from others or online is strictly prohibited. Electronic submission on Canvas is mandatory.

1. **Probability** (10 points) Suppose that we have three coloured boxes r (red), b (blue), and g (green). Box r contains 3 apples, 4 oranges, and 3 limes, box b contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probabilities $p(r) = 0.2$, $p(b) = 0.2$, $p(g) = 0.6$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

1.We need to find the probability of an apple i.e p(a). For this we apply the law of total probability:

$$
\begin{aligned}
p(a) &= p(a|b) * p(r) + p(a|b) * p(b) + p(a|g) * p(g) \\
p(a) &= .3 * .2 + .5 * .2 + .3 * .6 \\
p(a) &= .06 + .1 + .18 \\
p(a) &= .34
\end{aligned}
$$

2. Here we use Bayes rule to find the probability of $p(g|o)$

$$
\begin{aligned}
p(g|o) &= \frac{p(o|g) * p(g)}{p(o)} \\
p(o) &= p(o|b) * p(r) + p(o|b) * p(b) + p(o|g) * p(g) \\
&\text{Thus,} \\
p(g|o) &= \frac{0.3 * 0.6}{.4 * .2 + .1 + .3 * .6} \\
p(g|o) &= \frac{.18}{.08 + .1 + .18} \\
p(g|o) &= \frac{.18}{.36} \\
p(g|o) &= 0.5
\end{aligned}
$$

2. **Maximum Likelihood** (10 points) Assuming data points are independent and identically distributed (i.i.d.), the probability of the data set given parameters: μ and σ^2 (the likelihood function):

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

Please calculate the solution for μ and σ^2 using Maximum Likelihood (ML) estimator.

$$f(x_n|\mu,\sigma^2) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_n-\mu)^2}/(2\sigma^2)$$

To find the MLE we maximize the log likelikhood function to make the calculation easier as log will simplify the powers

$$\begin{aligned}
\log f(x_n|\mu,\sigma^2) &= \sum_{n=1}^{N} \log(\frac{1}{\sigma\sqrt{2\pi}} e^{-(x_n-\mu)^2}/(2\sigma^2)) \\
&= \sum_{n=1}^{N} [-\log(\sigma\sqrt{2\pi}) + \log(e^{-(x_n-\mu)^2}/(2\sigma^2))] \\
&= \sum_{n=1}^{N} [-\log(\sigma) - \log(\sqrt{(}2\pi) - ((x_n-\mu)^2)/(2\sigma^2)]
\end{aligned}$$

Now we partially differentiate the above eqaution wrt to σ and μ

1. σ:

$$\begin{aligned}
\frac{\partial LL(\mu,\sigma^2)}{\partial\sigma} &= \sum_{n=1}^{N} \frac{-1}{\sigma} + \frac{2(x_n-\mu)^2)}{2\sigma^3} \\
&= \frac{-n}{\sigma} + \sum_{n=1}^{N} \frac{(x_n-\mu)}{(\sigma^3)}
\end{aligned}$$

We set the above eqaution to zero and solve for σ^2 MLE:

$$\begin{aligned}
n\sigma^2 &= \sum_{n=1}^{N} (x_n-\mu)^2 \\
(\sigma^2)_{Mle} &= \frac{1}{n} \sum_{n=1}^{N} (x_n-\mu)^2
\end{aligned}$$

Simplifying this using μ_{mle} we get:

$$\sigma^2_{mle} = \frac{n}{n-1} S^2$$

Here S^2 is the sample variance and as we can see the sample variance is biased.

2.μ:

$$\frac{\partial LL(\mu, \sigma^2)}{\partial \sigma} = \sum_{n=1}^{N} \frac{2(x_n - \mu)}{2\sigma^2}$$

$$0 = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)$$

$$\mu_{mle} = \frac{1}{n} \sum_{n=1}^{N} x_n$$

From the above equation we can see that mean is unbiased and same as the sample mean.

3. **Maximum Likelihood** (15 points) We assume there is a true function $f(\mathbf{x})$ and the true value is given by $y = f(x) + \epsilon$ where ϵ is a Gaussian distribution with mean 0 and variance σ^2. Thus we can write:

$$p(y|x, w, \beta) = \mathcal{N}(y|f(x), \beta^{-1})$$

where $\beta^{-1} = \sigma^2$.

Assuming the data points are drawn independently from the distribution, we obtain the likelihood function:

$$p(\mathbf{y}|\mathbf{x}, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(x), \beta^{-1})$$

Please show that maximizing the likelihood function is equivalent to minimizing the sum-of-squares error function.

Here we want to show that maximizing likelihood is equivalent to minimizing the sum of of squares:

Let us consider a fucntion

$$y = wx_i + \epsilon$$

since the added gaussian noise is across throughout we can assume that the y will also be a gaussian

1. We write the log likelihood of the above function in gaussian form:

$$\mathcal{N}(y_i|x, w, \beta^{-1}) = \prod_{i=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(y_i - w^T x_i)^2}$$

We now take log on both sides to help make the calculation easier.

$$\ln(y|x, w, \beta^{)} = \frac{-\beta}{2} \sum_{i=1}^{N} (y_i - wx_i)^2 \frac{N}{2} \ln(\beta) - \frac{N}{2} \ln(2\pi)$$

We now maximize the log likelihood for the terms including w and ignore the rest

$$\hat{w} = \underset{w}{\text{argmax}}\, LL(w) = \underset{w}{\text{argmax}} -\frac{\beta}{2}\sum_{i=1}^{N}(y_i - wx_i)^2$$
$$= argmin_w - \frac{\beta}{2}\sum_{i=1}^{N}(y_i - wx_i)^2$$

The above equation shows that maximizing log likelihood is equivalent to minimizing the sum of sqaures function. Differentiating the above function w we get:

$$w_{\text{mle}} = (\sum_{i=1}^{N} x_i x_i^T)^{-1}\sum_{i=1}^{N} y_i x_i$$
$$= (X^T X)^{-1} X^T Y$$

As we can see the above equation is the closed form of the linear regression.

4. **MAP estimator** (20 points) Given input values $\mathbf{x} = (x_1, ..., x_N)^T$ and their corresponding target values $\mathbf{y} = (y_1, ..., y_N)^T$, we estimate the target by using function $f(x, \mathbf{w})$ which is a polynomial curve. Assuming the target variables are drawn from Gaussian distribution:

$$p(y|x, \mathbf{w}, \beta) = \mathcal{N}(y|f(x, \mathbf{w}), \beta^{-1})$$

and a prior Gaussian distribution for $\mathbf{w}$:

$$p(\mathbf{w}|\alpha) = (\frac{\alpha}{2\pi})^{(M+1)/2} \exp(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w})$$

Please prove that maximum posterior (MAP) is equivalent to minimizing the regularized sum-of-squares error function. Note that the posterior distribution of $\mathbf{w}$ is $p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \alpha, \beta)$. **Hint: use Bayes' theorem.** We assume D = (M+1)

Here we assume a 0 mean spherical gaussian prior given by:

$$p(w) = \mathcal{N}(0|\alpha^{-1}\mathbf{I}) = (\frac{\alpha}{2\pi})^D (e^{-\frac{\alpha}{2}w^T w})$$
$$= (\frac{\alpha}{2\pi})^D (e^{-\frac{\alpha}{2}||w||^2})$$

The posterior distribution can be expressed as:

$$= p(w|x, y, \alpha, \beta) \propto p(y|x, w, \beta)p(w|\alpha)$$
$$= \log p(w|x, y, \alpha, \beta) = \log p(y|x, w, \beta) + \log p(w|\alpha)$$
$$= \log p(w|x, y, \alpha, \beta) \propto -\frac{\beta}{2}\sum_{i=1}^{N}(y_i - wx_i)^2 - \frac{\alpha}{2}w^T w(\text{ignoring constants})$$
$$= \underset{w}{\text{argmax}} \log p(w|x, y, \alpha, \beta) = \underset{w}{\text{argmin}} \frac{\beta}{2}\sum_{i=1}^{N}(y_i - wx_i)^2 + \frac{\alpha}{2}w^T w$$

Differentiating w.r.t to w and equalling to 0 gives:

$$w_{\text{map}} = (X^T X + \frac{\alpha}{\beta}\mathbf{I})^{-1} X^T Y$$

As we can see this is equivalent to the closed form solution of ridge regression.

5. **Linear regression** (45 points) Please choose **one** of the below problems. You will need to **submit your code**.

   a) UCI Machine Learning: Conventional and Social Media Movies

   Please apply both Lasso regression and Ridge regression algorithms on this dataset for predicting movie ratings. You do not need to use all the features. Report how you divide data into training, validation, and testing. Run both regression algorithms 10 times and report the mean and standard deviation of the training error. Report the mean squared error (MSE) on the testing data.

   a) UCI Machine Learning: Residential Building Data Set Data Set

   Please apply both Lasso regression and Ridge regression algorithms on this dataset for predicting the house sale prices. You do not need to use all the features. Report how you divide data into training, validation, and testing. Run both regression algorithms 10 times and report the mean and standard deviation of the training error. Report the mean squared error (MSE) on the testing data.

   Uploaded as a python notebook along with this pdf.