Informal Notes for Mathematics for Machine Learning
Imperial College London

Lecture 5: Taking a Probabilistic Perspective on our Models

Lecturer: Matthew Wicker

1 Learning Objectives

In the first two weeks of course we have seen that by choosing our model (e.g., making some assumptions about input-output relationships) and choosing a loss function we can learn in both linear regression and neural network models. We have observed that we can solve directly for the "best" parameter or search for it by using automatic differentiation in order to compute numerical derivatives which can be used in gradient descent. However, there are several steps of this process that we have taken for granted that we hinted at when motivating this class: we have not yet seen *why* we made particular decisions. In particular, up until now the loss functions we have chosen have been prescribed to us. In this lecture, we will see how taking a probabilistic perspective on our models allows us to quickly identify a principled loss function for choosing the best parameter.

2 Probability Density Estimation

Thus far, we have not engaged in any meaningful way with probability. But this lecture will show us why it is very important to do so in order to understand core machine learning principles. Let us start by motivating the use of probability in density estimation. Recall from our discussion on problem settings that density estimation is an unsupervised learning task where we take a dataset $\{\boldsymbol{x}^{(i)}\}_{i=0}^{N}$ with our inputs $\boldsymbol{x} \in \mathbb{R}^n$ and we attempt to learn about the distribution from which are samples are drawn. This is a natural place to start coming from supervised learning as we can follow the same sort of framework:

1. Make an assumption about the distribution the data is coming from. $x \sim p_\theta$
2. Identify the likelihood model for our data (and the negative log likelihood, typically) .
3. Solve for the best parameter.

This is of course very similar to the process we looked at in our supervised learning set ups with one key exception: in those settings you were asked to (sort of) arbitrarily choose a loss function and then solve. But here, we are doing something more probabilistically principlled as we will see now for the Bernoulli distribution.

2.1 MLE in Unconditional Density Estimation

Just so that this section is self-contained, recall that (unconditional) density estimation observes a set of samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^N$ where $\boldsymbol{x}^{(i)} \sim \pi(x)$. Our goal is then to perform what can be considered *unconditional* density estimation where we try to fit a parameterized distribution $p(x|\theta)$, sometimes written $p_\theta(x)$, such that this distribution is as close to the unknown distribution as possible. That is we want to find some parameterization such that $p(\boldsymbol{x}|\theta) \approx \pi(x)$. As a simple starting example, let us consider one of the most classical statistical examples: a biased coin.

2.1.1 MLE for Bernoulli Density

Assume we have a biased coin and we would like to estimate how frequently heads comes up. In this case, we have flipped the coin N times making up our dataset with $x \in \{0, 1\}$ (we no longer denote this as a vector since it is an integer). Given the binary sample space, a natural choice is to model this process as a Bernoulli random variable:

$$p_\theta(X = x) = \begin{cases} \theta & \text{if } x = 1 \\ (1-\theta) & \text{if } x = 0 \end{cases}$$

To start, let us model the likelihood of observing our given dataset for an arbitrary parameter θ:

$$\mathcal{L}(\theta) = \prod_{i=1}^N P(X = x^{(i)}) = \prod_{i=1}^N \theta^{x^{(i)}}(1-\theta)^{1-x^{(i)}}$$

A common theme that we will see throughout the course is that when dealing with likelihoods, it will prove both mathematically and numerically convenient for our analysis to take the log of the likelihood:

$$\log \mathcal{L}(\theta) = \sum_{i=1}^N \left(x^{(i)} \log \theta + (1 - x^{(i)}) \log(1-\theta)\right)$$

To find the MLE, as we did for our linear regression model, we differentiate the log likelihood with respect to θ and set it equal to 0:

$$\frac{d}{d\theta}(\log \mathcal{L}(\theta)) = \sum_{i=1}^N \left(\frac{x^{(\theta)}}{\theta} - \frac{1-x^{(i)}}{1-\theta}\right) = 0$$

Multiplying through by $\theta(1-\theta)$ and introducing the assumption that $\theta \neq 0$ and $\theta \neq 1$:

$$\sum_{i=1}^N \left(x^{(i)}(1-\theta) - (1-x^{(i)})\theta\right) = 0$$

With elementary algebra (simply expand and rearrange into two sums, we get that)

$$\sum_{i=1}^N x^{(i)} = N\theta \implies \theta = \frac{1}{N}\sum_{i=1}^N x^{(i)}$$

So, the MLE for p, the probability of success in a Bernoulli trial, is the sample mean of the observed outcomes. This points to one nice property of the MLE, which is that it is typically interpretable.[1] Now, critically, we can answer questions like "what is the coin most likely next outcome, heads or tails?" However, what we would like to emphasize in this process is the critical steps of deriving the MLE. Namely (1) Write out the probabilistic model, (2) Identify the likelihood, and (3) set the gradient of the log to 0 and solve. In the exercises at the end of Lecture 1 you were asked to compute the MLE for the Gaussian distribution mean. I would urge students who have not taken a statistics course before to go back to that exercise.

2.2 MLE in Conditional Density Estimation

Let us now turn to applying the idea of maximizing the likelihood of the observed data to our supervised learning set up with, $\mathcal{D} := \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ in our linear regression set up. Now, one key benefit on introducing probability in learning is the ability to account for observational noise, something we have ignored up to this point. Taking into account observational noise (e.g., from sensors that we get our measure ments from) we have that our model is now:

$$\hat{y}^{(i)} = \boldsymbol{x}^{(i)\top}\theta + \epsilon^{(i)}$$

where $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$. We will make the assumption that the ϵ_i for $i = 1, \ldots, N$ are all independent random variables. Now, let us write the likelihood of observing the data mathematically. Since we only model the outputs $y^{(1)}, \ldots, y^{(N)}$ probabilistically, we have

$$p(y^{(1)}, \ldots, y^{(N)} | \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}, \theta, \sigma) = \prod_{i=1}^{N} p(y^{(i)} | \boldsymbol{x}^{(i)}, \theta, \sigma^2) \text{ (using independence)}$$

Returning to our first lecture on independent and identically distributed notice how in this probabilistic formulation of the linear model we have made use of both of these assumptions. Expanding our ϵ term into its probability density, we can see that our model return parameters according to $\hat{y}^{(i)} = \boldsymbol{x}^{(i)\top}\theta + \mathcal{N}(0, \sigma^2)$. Given that we are adding a deterministic quantity to the mean of our Gaussian distribution, we can re-write the model as, $\hat{y}^{(i)} = \mathcal{N}(\boldsymbol{x}^{(i)\top}\theta, \sigma^2)$. This makes the probabilistic nature of our model much more clear and allos us now to expand out the above liklihood using our knowledge of the analytical form of the Gaussian distribution:

$$p(y^{(1)}, \ldots, y^{(N)} | x^{(1)}, \ldots, x^{(N)}, \theta, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \boldsymbol{x}^{(i)\top}\theta)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y^{(i)} - \boldsymbol{x}^{(i)\top}\theta)^2\right)$$

[1] We break down further pros and cons in the lecture slides

$$\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \boldsymbol{x}^{(i)\top}\theta)^2}{2\sigma^2}\right)$$

We want to find parameters θ and σ that maximize the likelihood. Since the logarithm, $\log : \mathbb{R}^+ \to \mathbb{R}$, is a monotonic, increasing function, we can instead maximize the log-likelihood which we can Now, expressed in a slightly more convenient form, the log-likelihood can be written as:

$$LL(y^{(1)}, \ldots, y^{(N)} | \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}, \theta, \sigma) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y^{(i)} - \boldsymbol{x}^{(i)\top}\theta)^2$$

We can express everything in vector notation as we have done in all previous analysis with the linear regression model:

$$LL(y|X, \theta, \sigma) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\boldsymbol{X}\theta - \boldsymbol{y})^T (\boldsymbol{X}\theta - \boldsymbol{y})$$

Of course maximizing a function and minimizing its negation yeild identical argmaxes, so we can minimize the negative log-likelihood:

$$NLL(\boldsymbol{y}|\boldsymbol{X}, \theta, \sigma^2) = \frac{1}{2\sigma^2} (\boldsymbol{X}\theta - \boldsymbol{y})^\top (\boldsymbol{X}\theta - \boldsymbol{y}) + \frac{N}{2} \log(2\pi\sigma^2)$$

Now, let us take a look at the expanded form above. We have seen a function that is almost identical to this when deriving our least squares estimate for linear regression. The critical difference here and in the linear regression case is that here we did not make the arbitrary decision to adopt the mean squared error. Instead we simply formulated a probabilistic model. Recall the mean squarred error was stated as:

$$\mathcal{L}(\theta) = \frac{1}{2N} (\boldsymbol{X}\theta - \boldsymbol{y})^\top (\boldsymbol{X}\theta - \boldsymbol{y})$$

And when minimizing with respect to θ, the two objectives $\mathcal{L}(\theta)$ and $NLL(\theta)$ are the same up to a constant additive and multiplicative factor. Thus, we know that the maximum likelihood estimate for θ is given by:

$$\theta_{\text{MLE}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

We can also find the maximum likelihood estimate for σ.

$$\sigma^2_{\text{MLE}} = \frac{1}{N} (\boldsymbol{X}\theta_{\text{MLE}} - \boldsymbol{y})^\top (\boldsymbol{X}\theta_{\text{MLE}} - \boldsymbol{y})$$

In lecture we will also derive the cross-entropy loss and formulate the logistic regression model.

2.3 Probabilistic modelling connections to information theory

In the above, we saw how formulating our linear regression model probabilistically allowed informed a principled selection of the MSE loss function. Here we provide another key strength of probabilistic modelling which is that it allows us to connect our ML problems to rich fields of mathematics whose tools might prove very useful when analyzing the performance of our models. One such field is that of information theory. A central component of information theory is the entropy of a distribution:

Definition 2.1 (Entropy of a Distribution)**.** The entropy of a probability distribution p with respect to a discrete random variable X is defined as:

$$H(X) = -\sum_x p(x) \log(p(x))$$

where: $H(X)$ is the entropy of the distribution, x represents possible values of the random variable X, and $p(x)$ is the probability mass function associated with X.

While entropy is the cornerstone of information theory, it is also used widely in machine learning to quantify uncertainty of predictions. Consider the entropy of our Bernoulli distribution. It has a single peak at 0.5 which is when one might say they are the most uncertain about the next outcome of from the flip of a coin. A key function in information theory is the KL divergence:

Definition 2.2 (Kullback-Leibler (KL) Divergence)**.** The Kullback-Leibler (KL) divergence between two probability distributions p and q with respect to a discrete random variable X is defined as:

$$\mathrm{KL}(p \,||\, q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) = \sum_x p(x)\log(p(x)) - \sum_x p(x)\log(q(x))$$

where: $\mathrm{KL}(p \,||\, q)$ is the KL divergence, x represents possible values of the random variable X, $p(x)$ and $q(x)$ are the probability mass functions associated with X for distributions p and q, respectively.

We will dive into the mathematics and workings of the KL divergence in more detail if and when we discuss approximate inference; however, I want to now simply connect the maximum likelihood framework that we have observed to these two information theoretic quantities. Assume we are fitting a distribution q to a set of observations. We have that:

$$\begin{aligned}
\theta_{\text{MLE}} &= \arg\max_\theta \prod_{i=1}^N q(x^{(i)}|\theta) \quad \text{(Maximum Likelihood)} \\
&= \arg\max_\theta \sum_{i=1}^N \log(q(x^{(i)}|\theta)) \\
&= \arg\min_\theta \frac{1}{N}\sum_{i=1}^N \log(q(x^{(i)}|\theta)) - \frac{1}{N}\sum_{i=1}^N p(x^{(i)}) \\
&= \arg\min_\theta \frac{1}{N}\sum_{i=1}^N \log\left(\frac{p(x^{(i)})}{q(x^{(i)}|\theta)}\right) \\
&\to \arg\min_\theta \mathrm{KL}(p \,||\, q_\theta)
\end{aligned}$$

where the third step is simply replacing the sum with an average and adding a term that in no way depends on θ, thus will not affect the location of the argmax/argmin parameter. What we see here is that by formulating our model probabilistically, we can see maximum likelihood estimation as simply a value that minimizes the KL divergence. From here, we now have in roads to analyze our algorithms with information theory.

2.4 MAP in Conditional Density Estimation

Returning now to our supervised learning scenario, lets see how we can take our probabilistic modelling to the next level. Above we established that what we want to do is find a good probabilistic model, $p(y|\boldsymbol{x}, \theta)$. Recall from our earlier lectures that we introduced basis expansion as a critical tool to increase the strength of our linear-in-the-parameters model:

$$y = \phi(\boldsymbol{x})\theta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

we covered that one potential choice for ϕ allows us to fit arbitrarily high degree polynomials to our data. At first glance, this is great. It increases the strength of our model which allows us to capture more complex relationships between the input features and labels. This increase in model power is a double edged sword, however. It is true that by using a degree $n + 1$ polynomial, we interpolate any n data points we wish. We will cover exactly why this kind of over-fitting is catastrophically bad in Lecture 8 (and indeed you may have seen this in previous courses). However, as we discussed in lecture, one major drawback to the MLE is that it returns only a point estimate with no ability to quantify uncertainty which can lead to over-fitting. One principled way to try to counter-act over-fitting is by ensuring that the value of our model parameters remain small in magnitude. One way to do this is to set a *prior distribution* over our parameters. Thus, the objective we want to minimize is:

$$\theta_{\text{MAP}} = \operatorname*{argmax}_{\theta} \mathbb{E}_{p(\boldsymbol{x},y)} p(y|\boldsymbol{x}, \theta) p(\theta) \tag{1}$$

$$= \operatorname*{argmax}_{\theta} \mathbb{E}_{p(\boldsymbol{x},y)} p(y|\boldsymbol{x}, \theta) \mathcal{N}(\theta; 0, \tau^2 \mathbb{I}) \tag{2}$$

Where the latter equation fills in that our prior distribution over our parameters is $\mathcal{N}(0, \tau^2\mathbb{I})$. Now assuming observational noise as well, lets try to compute this minimization of this objective:

$$p(y|\boldsymbol{x}, \theta) = p(y|\boldsymbol{x}, \theta)\mathcal{N}(\theta; 0, \tau^2\mathbb{I}) \tag{3}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(- \frac{(\boldsymbol{y} - \boldsymbol{X}\theta)^\top (\boldsymbol{y} - \boldsymbol{X}\theta)}{2\sigma^2} \Big) \mathcal{N}(\theta; 0, \tau^2 \mathbf{I}) \tag{4}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(- \frac{(\boldsymbol{y} - \boldsymbol{X}\theta)^\top (\boldsymbol{y} - \boldsymbol{X}\theta)}{2\sigma^2} \Big) \cdot \frac{1}{(2\pi\tau^2)^{d/2}} \exp\Big(- \frac{\theta^\top \theta}{2\tau^2} \Big) \tag{5}$$

While this looks quite scary, a first simplifying step is to notice that the maximum that we seek is not going to be affected by multiplying by constants. Thus we can drop the leading fractions to give us:

$$\operatorname*{argmax}_{\theta} \exp\Big(- \frac{(\boldsymbol{y} - \boldsymbol{X}\theta)^\top (\boldsymbol{y} - \boldsymbol{X}\theta)}{2\sigma^2} \Big) \cdot \exp\Big(- \frac{\theta^\top \theta}{2\tau^2} \Big) \tag{6}$$

$$= \operatorname*{argmax}_{\theta} \exp\Big(- \frac{(\boldsymbol{y} - \boldsymbol{X}\theta)^\top (\boldsymbol{y} - \boldsymbol{X}\theta)}{2\sigma^2} - \frac{\theta^\top \theta}{2\tau^2} \Big) \tag{7}$$

max: -ve min

Using the fact that log is monotone and converting argmax to argmin by flipping signs, we get that the above is equal to:

$$= \underset{\theta}{\operatorname{argmin}} \exp\Big(\frac{(\boldsymbol{y} - \boldsymbol{X}\theta)^\top(\boldsymbol{y} - \boldsymbol{X}\theta)}{2\sigma^2} + \frac{\theta^\top\theta}{2\tau^2}\Big) \qquad (8)$$

$$= \underset{\theta}{\operatorname{argmin}} \exp\Big((\boldsymbol{y} - \boldsymbol{X}\theta)^\top(\boldsymbol{y} - \boldsymbol{X}\theta) + \frac{\sigma^2}{\tau^2}\theta^\top\theta\Big) \qquad (9)$$

As an exercises, follow along with the steps from above to find the minimum of this function (hint: we did this at the start of our last lecture minus the additional constant). You should arrive at:

$$\theta^{\text{MAP}} = \big(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}_d\big)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

Notice, this is different than what we obtained in Lecture 4. Notice that as τ tends to infinity, we get that the resulting inverse leaves us with $\theta^{\text{MAP}} = \mathbf{0}$. Obviously, this extreme is as undesirable as the case where we over-fit. In practice, one selects a *happy medium* using mathematical principles that we will derive once we have a few more probabability tools under our belts.

A Lecture 4: Probabilistic Modelling Principles

Question 1 (Training translation models)**.** Imagine you want to train a neural network $T_\theta(\cdot)$ to translate French words to English words. Assume you are given a dataset $\mathcal{D} = \{(f_n, e_n)\}_{n=1}^N$ where f_n is a French word and e_n is an English word. Suppose the vocabulary of French and English is $\mathcal{F}$ and $\mathcal{E}$, respectively.

a. Assuming a probabilistic model $p(e|T_\theta(f))$, which distribution would you choose for this model?

b. Continuing a), what is the corresponding MLE objective?

Question 2 (Clustering)**.** We consider a clustering task where given a dataset $\mathcal{D} = \{x^{(1)}, ..., x^{(N)}\}$, we would like to group them into K clusters. The model we will use here is a Gaussian mixture model:

$$\text{GMM:} \quad p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \sigma^2), \quad \theta = \{\pi_k, \mu_k, \}_{k=1}^K, \sigma^2.$$

a. What is the MLE objective for this clustering task?

b. Derive the gradient of the MLE objective w.r.t. μ_k. What is the fixed-point equation for finding the optimal $\{\mu_k\}$ parameters?

Question 3 (Geometric interpretation of linear regression)**.** Consider the following linear regression model:

$$y = \theta^\top \phi(\boldsymbol{x}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

For a given dataset $\{(\boldsymbol{x}^{(N)}, y^{(N)})\}_{n=1}^N$, Writing $\boldsymbol{\Phi} = (\phi(\boldsymbol{x}^{(1)}), \phi(\boldsymbol{x}_2), ..., \phi(\boldsymbol{x}^{(N)}))^\top$ and $\boldsymbol{y} = (y^{(1)}, ..., y^{(N)})^\top$, we have the optimal solution satisfies $\theta^* = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}$. Show that by using the optimal parameter θ^*, the prediction $\hat{\boldsymbol{y}} = (\hat{y}_1, ..., \hat{y}_N), \hat{y}_n = (\theta^*)^\top \phi(\boldsymbol{x}^{(N)})$ is the projection of $\boldsymbol{y}$ onto the sub-space spanned by the columns of Φ.

(Hint: consider singular value decomposition.)

Question 1 (Training translation models)**.** Imagine you want to train a neural network $T_\theta(\cdot)$ to translate French words to English words. Assume you are given a dataset $\mathcal{D} = \{(f_n, e_n)\}_{n=1}^N$ where f_n is a French word and e_n is an English word. Suppose the vocabulary of French and English is $\mathcal{F}$ and $\mathcal{E}$, respectively.

a. Assuming a probabilistic model $p(e|T_\theta(f))$, which distribution would you choose for this model?

b. Continuing a), what is the corresponding MLE objective?

a) Categorical dist. $T_\theta(\cdot) : F \to \mathbb{R}^{|\mathcal{E}|}$ - map french word to real val vector length $|\mathcal{E}|$

$p(e|T_\theta(f)) :$ Categorical (Softmax $(T_\theta(f))$) -> max prob of getting correct word

b) $p(e_n | T_\theta(f))$. Softmax: $S(x_i)$: $\frac{e^{x_i}}{\sum_{j=1}^{\hat{}} e^{x_j}}$ -> $S(T_\theta(f_i)) : \frac{e^{T_\theta(f_i)|e_i}}{\sum_{j=1}^{|\mathcal{E}|} e^{T_\theta(f_i)|e_j}}$ ← pick out correct word

$L(\theta)$: log likelihood : $\log p(e_1|T_\theta(f_1)) \to \log \sum_{n=1}^{N} p(e_n|T_\theta(f_n))$

θ: $\underset{\theta}{\operatorname{argmax}} L(\theta) = \underset{\theta}{\operatorname{argmax}} \log \sum_{i=1}^{\tilde{}} \frac{e^{T_\theta(f_i)|e_i}}{\sum_{j=1}^{|\mathcal{E}|} e^{T_\theta(f_i)|e_j}}$

Question 2 (Clustering). We consider a clustering task where given a dataset $\mathcal{D} = \{x^{(1)}, ..., x^{(N)}\}$, we would like to group them into K clusters. The model we will use here is a Gaussian mixture model:

$$\text{GMM:} \quad p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \sigma^2), \quad \theta = \{\pi_k, \mu_k, \}_{k=1}^{K}, \sigma^2.$$

a. What is the MLE objective for this clustering task?

b. Derive the gradient of the MLE objective w.r.t. μ_k. What is the fixed-point equation for finding the optimal $\{\mu_k\}$ parameters?

a) $\mathcal{N}(x; \mu_k, \sigma^2) \rightarrow \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

likelihood: $p(D|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta) \quad \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \sigma^2), \quad \theta = \{\pi_k, \mu_k\}_{k=1}^{K}, \sigma^2$

Take log: $L(\theta) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \sigma^2)\right)$

b) diff $L(\theta)$ w.r.t μ_k

$\mathcal{N}(x; \mu_k, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \rightarrow \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \times \frac{d}{d\mu}\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$

$x^2 - 2x\mu + \mu^2$
$-2x + 2\mu$

$\frac{d}{d\mu}$

$-\frac{1}{2\sigma^2} \times 2(x-\mu) \times -1$

$= \frac{x-\mu}{\sigma^2}$

$\frac{d}{d\mu}(\mathcal{N}(x^{(i)}; \mu_k, \sigma^2)) = \mathcal{N}(x^{(i)}; \mu_k, \sigma^2) \frac{x^{(i)} - \mu_k}{\sigma^2}$

$\frac{\partial L(\theta)}{\partial \mu} = \frac{1}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}; \mu_k, \sigma^2)} \pi_k \frac{x^{(i)} - \mu_k}{\sigma^2} \mathcal{N}(x, \mu_k, \sigma^2)$

$= \frac{\pi_k (x - \mu_k) \mathcal{N}(x; \mu_k, \sigma^2)}{\sigma^2 \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \sigma^2)} = 0$

Bayes: $p(k|x) = \frac{p(x|k)p(k)}{p(x)} = \frac{\mathcal{N}(x; \mu_k, \sigma^2)\pi_k}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \sigma^2)}$

$\therefore D_\mu L(\theta) = p(k|x) \frac{x - \mu_k}{\sigma^2} = 0$

$p(k|x)x - p(k|x)\mu_k = 0$

$\mu_k = \frac{p(k|x)x}{p(k|x)}$

Question 3 (Geometric interpretation of linear regression)**.** Consider the following linear regression model:

$$y = \theta^\top \phi(\boldsymbol{x}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

For a given dataset $\{(\boldsymbol{x}^{(N)}, y^{(N)})\}_{n=1}^N$, Writing $\boldsymbol{\Phi} = (\phi(\boldsymbol{x}^{(1)}), \phi(\boldsymbol{x}_2), ..., \phi(\boldsymbol{x}^{(N)}))^\top$ and $\boldsymbol{y} = (y^{(1)}, ..., y^{(N)})^\top$, we have the optimal solution satisfies $\theta^* = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}$. Show that by using the optimal parameter θ^*, the prediction $\hat{\boldsymbol{y}} = (\hat{y}_1, ..., \hat{y}_N), \hat{y}_n = (\theta^*)^\top \phi(\boldsymbol{x}^{(N)})$ is the projection of $\boldsymbol{y}$ onto the sub-space spanned by the columns of Φ.

(Hint: consider singular value decomposition.)