CS 6190: Probabilistic Machine Learning Spring 2023

Homework 1

Handed out: 2 Feb, 2023
Due: 11:59pm, 20 Feb, 2023

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down your own solution. Please keep the class collaboration policy in mind.
- Feel free discuss the homework with the instructor or the TAs.
- Your written solutions should be brief and clear. You need to show your work, not just the final answer, but you do *not* need to write it in gory detail. Your assignment should be **no more than 20 pages**. Every extra page will cost a point.
- Handwritten solutions will not be accepted.
- The homework is due by **midnight of the due date**. Please submit the homework on Canvas.

Analytical problems [80 points + 30 bonus]

1. [8 points] A random vector, $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$ follows a multivariate Gaussian distribution,

$$p(\mathbf{x}) = \mathcal{N}(\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} | \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}).$$

Show that the marginal distribution of $\mathbf{x}_1$ is $p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

solution

$$Q(x_1, x_2) = [(x_1 - \mu_1)^T, (x_2 - \mu_2)^T] \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} [\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}]$$

Assuming that $\boldsymbol{\Sigma}_{ij}^{-1} = \boldsymbol{\Sigma}^{ij}$

$$\boldsymbol{\Sigma}^{11} = \sum_{11}^{-1} + \sum_{11}^{-1} \sum_{12} (\sum_{22} - A_{12}^T \sum_{11}^{-1} \sum_{12})^{-1} \sum_{12}^T \sum_{11}^{-1}$$

$$\boldsymbol{\Sigma}^{22} = \sum_{22}^{-1} + \sum_{22}^{-1} \sum_{12}^T (\sum_{11} - \sum_{12} \sum_{22}^{-1} \sum_{12}^T)^{-1} \sum_{12} \sum_{22}^{-1}$$

$$\boldsymbol{\Sigma}^{12} = (\boldsymbol{\Sigma}^{21})^T = -\sum_{11}^{-1} \sum_{12} (\sum_{22} - \sum_{12}^T \sum_{11}^{-1} \sum_{12})^{-1}$$

So then, subsituting those into $Q(x_1, x_2)$ we get

$$Q(x_1, x_2) = (x_1 - \mu_1)^T (\sum_{11}^{-1} + \sum_{11}^{-1} \sum_{12} (\sum_{22} - A_{12}^T \sum_{11}^{-1} \sum_{12})^{-1} \sum_{12}^T \sum_{11}^{-1})(x_1 - \mu_1) - 2(x_1 - \mu_1)^T (\sum_{11}^{-1} \sum_{12} (\sum_{22} - \sum_{12}^T \sum_{11}^{-1} \sum_{12})^{-1})(x_2 - \mu_2) + (x_2 - \mu_2)^T ((\sum_{22} - \sum_{12}^T \sum_{11}^{-1} \sum_{12})^{-1})(x_2 - \mu_2)$$

$$= (x_1 - \mu_1)^T \sum_{11}^{-1} (x_1 - \mu_1) + (x_1 - \mu_1)^T \sum_{11}^{-1} \sum_{12} (\sum_{22} - A_{12}^T \sum_{11}^{-1} \sum_{12})^{-1} \sum_{12}^T \sum_{11}^{-1})(x_1 - \mu_1) - 2(x_1 - \mu_1)^T (\sum_{11}^{-1} \sum_{12} (\sum_{22} - \sum_{12}^T \sum_{11}^{-1} \sum_{12})^{-1})(x_2 - \mu_2) + (x_2 - \mu_2)^T ((\sum_{22} - \sum_{12}^T \sum_{11}^{-1} \sum_{12})^{-1})(x_2 - \mu_2)$$

$$= (x_1 - \mu_1)^T \sum_{11}^{-1} (x_1 - \mu_1) + ((x_2 - \mu_2) - \sum_{12}^T \sum 11^{-1} (x_1 - \mu_1))^T (\sum_{22} - \sum_{12}^T \sum_{11}^{-1} \sum_{12})^{-1} ((x_2 - \mu_2) - \sum_{12}^T \sum_{11}^{-1} (x_1 - \mu_1))$$

$$A = \sum_{22} - \sum_{12}^T \sum_{11}^{-1} \sum_{12}$$

$$p(x) = \frac{1}{(2\pi)^{(n/2) | \sum |^{1/2}}} exp((-1/2) Q(x_1, x_2))$$

$$p(x_1) = \frac{1}{(2\pi)^{(n/2) | \sum_{11} |^{1/2}}} exp((-1/2)(x_1 - \mu_1)^T \sum_{11}^{-1} (x_1 - \mu_1)) = \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

2. [**Bonus**][10 points] Given a Gaussian random vector, $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We have a linear transformation, $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} + \mathbf{z}$, where $\mathbf{A}$ and $\mathbf{b}$ are constants, $\mathbf{z}$ is another Gaussian random vector independent to $\mathbf{x}$, $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \boldsymbol{\Lambda})$. Show $\mathbf{y}$ follows Gaussian distribution as well, and derive its form. Hint: using characteristic function. You need to check the materials by yourself.

3. [8 points] Show the differential entropy of the a multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\mathrm{H}[\mathbf{x}] = \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{d}{2}(1 + \log 2\pi)$$

where d is the dimension of $\mathbf{x}$.

solution

Differential entropy $= h(x) = -\int_X p(x) log_b p(x) dx$

in natural units: $= -E[log(p(x))]$

pdf of multivariate gaussian distribution: $\frac{1}{\sqrt{2\pi^d|\sum|}} exp(-1/2(x-\mu)^T \sum^{-1}(x-\mu))$

plug into natural units of differential entropy =

$-E[log(\frac{1}{\sqrt{2\pi^d|\sum|}} exp(-1/2(x-\mu)^T \sum^{-1}(x-\mu)))]$

take the log $= -E[-d/2 * log(2\pi) - 1/2 * log|\sum| - 1/2(x-\mu)^T \sum^{-1}(x-\mu)]$

$= d/2 * log(2\pi) + 1/2 * log|\sum| + 1/2E((x-\mu)^T \sum^{-1}(x-\mu))$

where $E((x-\mu)^T \sum^{-1}(x-\mu))$ can be simplified as follows:

$= E(tr((x-\mu)^T \sum^{-1}(x-\mu)))$

$= tr(\sum^{-1} E((x-\mu)(x-\mu)^T))$

$= tr(\sum^{-1} \sum)$

$= tr(I_d) = d$

plug into $H(x) = d/2 * log(2\pi) + 1/2 * log|\sum| + 1/2E((x-\mu)^T \sum^{-1}(x-\mu))$

$H(x) = d/2 * log(2\pi) + 1/2 * log|\sum| + (1/2)d$

$= d/2(log(2\pi) + 1) + 1/2log|\sum|$

4. [8 points] Derive the Kullback-Leibler divergence between two Gaussian distributions, $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \Lambda)$, i.e., $\mathrm{KL}(q||p)$.

$\mathrm{KL}(q||p) = \sum q(x) log(\frac{q(x)}{p(x)})$

for continuous distributions such as the normal distribution:

$= \int q(x) log(\frac{q(x)}{p(x)})$

$= -\int q(x) log(p(x)) dx + \int q(x) log(q(x)) dx$

$\int q(x) log(q(x)) dx = -1/2(1 + log(2\pi\Lambda^2))$

$-\int q(x) log(p(x)) dx = -\int q(x) log(\frac{1}{(2\pi\sum^2)^{0.5}}) * exp(\frac{-(x-m)^2}{2\sum^2}) dx$

$= 1/2 * log(2\pi\sum^2) - \int q(x) log(exp(\frac{-(x-m)^2}{2\sum^2}) dx$

$= 1/2 * log(2\pi\sum^2) - \int q(x)(\frac{-(x-m)^2}{2\sum^2}) dx$

$= 1/2 * log(2\pi\sum^2) + \frac{\int q(x)x^2 dx - \int q(x)2xm dx + \int q(x)m^2 dx}{2\sum^2}$

$\int q(x)y = var(y)$, and since $var(x^2) = \Lambda^2 + \mu^2$ we can say that

$= 1/2 * log(2\pi\sum^2) + \frac{\Lambda^2 + \mu^2 - 2\mu m + m^2}{2\sum^2}$

So, when we put it all together with q(x) and p(x) we get

$\mathrm{KL}(q||p) = 1/2 * log(2\pi\sum^2) + \frac{\Lambda^2 + \mu^2 - 2\mu m + m^2}{2\sum^2} - 1/2(1 + log(2\pi\Lambda^2))$

5. [8 points] Given a distribution in the exponential family,

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right).$$

Show that

$$\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} = \mathrm{cov}(\mathbf{u}(\mathbf{x})),$$

where cov is the covariance matrix.

$$\frac{d}{dn}log(Z(n)) = \frac{d}{dn}(log(\textstyle\sum_x)exp(Z(n)^T u(x)))$$
$$= exp(-Z(n))\textstyle\sum u(x)exp(Z(n)^T u(x))$$
$$\frac{d^2}{d^2n}log(Z(n)) = \frac{d}{dn}(exp(-Z(n))\textstyle\sum u(x)exp(Z(n)^T u(x)))$$
$$= exp(-Z(n))\textstyle\sum_x u(x)^2 exp(Z^T u(x)) - exp(-2Z(n))(\sum_x u(x)exp(Z^T u(x)))^2$$
$$= cov(u(x))$$

6. [4 points] Is $\log Z(\boldsymbol{\eta})$ convex or nonconvex? Why?

We know that the normalizer of the exponential family equation is convex, and it is shown how that is true on page 3 of this source. It shows how the second derivative of the normalizer is positive semi-definite, meaning that the normalizer is convex.:
https://www.stat.cmu.edu/ larry/=stat705/Lecture12a.pdf
The log of the normalizer is convex because we know that the normalizer is convex, and the log of a function whose second derivative is positive semi-definite means that the second derivative of the log of the function is positive semi-definite, making it convex.

7. [8 points] Given two random variables $\mathbf{x}$ and $\mathbf{y}$, show that

$$I(\mathbf{x}, \mathbf{y}) = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}]$$

where $I(\cdot, \cdot)$ is the mutual information and $H[\cdot]$ the entropy.

solution

$$I(x,y) = \textstyle\sum_{(x\epsilon X, y\epsilon Y)} p_{(X,Y)}(x,y)log(\frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)})$$
$$= \textstyle\sum_{(x\epsilon X, y\epsilon Y)} p_{(X,Y)}(x,y)log(\frac{p_{(X,Y)}(x,y)}{p_X(x)}) - \sum_{(x\epsilon X, y\epsilon Y)} p_{(X,Y)}(x,y)log(p_Y(y))$$
$$= \textstyle\sum_{(x\epsilon X, y\epsilon Y)} (p_X(x)p_{(Y|X=x)}(y))log(p_{(Y|X=x)}(y)) - \sum_{(x\epsilon X, y\epsilon Y)} p_{(X,Y)}(x,y)log(p_Y(y))$$
$$= -\textstyle\sum_{x\epsilon X} p_X(x)H(Y|X=x) - \sum_{y\epsilon Y} p_Y(y)logp_Y(y)$$
$$= -H(Y|X) + H(Y) = H(Y) - H(Y|X)$$

and since mutual information is symmetric, we know that I(x,y) = I(y,x), and so

$$H(Y) - H(Y|X) = H(X) - H(X|Y)$$

8. [24 points] Convert the following distributions into the form of the exponential-family distribution. Please give the mapping from the expectation parameters to the natural parameters, and also represent the log normalizer as a function of the natural parameters.

- Dirichlet distribution
- Gamma distribution
- Wishart distribution

solution

Exponential family distribution form = $p(x|\theta) = \frac{1}{z(\theta)} * h(x) * exp(dot(t(x), \theta))$

Dirichlet distribution

Dirichlet distribution is written as $Dir(x|\theta) = \frac{\Gamma(\sum_k \theta_k)}{\Pi_k \Gamma(\theta_k)} \Pi_k x_k^{\theta_k - 1}$

This can be re-written into the exponential family as

$$exp(\textstyle\sum_k (\theta_k - 1)log(x_k) - [\sum_k log(\Gamma(\theta_k)) - log(\Gamma(\sum_k \theta_k))])$$
$$Z(n) = [\textstyle\sum_k log(\Gamma(n_k)) - log(\Gamma(\sum_k n_k))]$$

$$\text{n} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

$$\text{T(x)} = \begin{bmatrix} \log(x_1) \\ \vdots \\ \log(x_n) \end{bmatrix}$$

$$h(x) = \frac{1}{\Pi_{i=1}^{k} x_i}$$

Gamma distribution

Gamma distribution is written as $p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} exp(-\beta x)$
putting the $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}$ into the exponential, we get

$$= exp(-\beta x) exp((\alpha - 1) log(x) - [log(\Gamma(\alpha) - \alpha log(\beta)]))$$

$$\text{n} = \begin{bmatrix} \alpha - 1 \\ \text{-}\beta \end{bmatrix}$$

$$h(x) = 1$$

$$\text{T(x)} = \begin{bmatrix} \text{log(x)} \\ \text{x} \end{bmatrix}$$

$$Z(n) = log(\Gamma(n_1 + 1)) - (n_1 + 1) log(-n_2)$$

Wishart distribution

this can be written as $f(x|V, n) = \frac{exp(-1/2tr(V^{-1}x)) det(x)^{(n-p-1)/2}}{2^{np/2} det(V)^{n/2} \Gamma_p(n/2)}$
in exponential family form we have:

$$\frac{exp(-1/2tr(V^{-1}x) + \frac{n-p-1}{2} log(|x|)}{2^{pn/2} |V|^{n/2} \Gamma_p(n/2)}$$

$$\text{n} = \begin{bmatrix} \text{(-1/2)V}^{-1} \\ \text{n - p - 1 } \frac{}{2} \end{bmatrix}$$

$$\text{h(x) = 1}$$

$$\text{T(x)} = \begin{bmatrix} \text{x} \\ \text{log—x—} \end{bmatrix}$$

$$Z(n) = -\frac{n}{2} log| - n_1| + log(\Gamma_p)(n/2)$$

9. [6 points] Does student t distribution (including both the scalar and vector cases) belong to the exponential family? Why?

No, the student t distribution does not belong to the exponential family because the exponential family is defined by a probability density function that should be able to be expressed in terms of a natural parameter and a normalizing constant, whereas the student t distribution does not have a natural parameter.

10. [6 points] Does the mixture of Gaussian distribution belong to the exponential family? Why?

$$p(\mathbf{x}) = \frac{1}{2}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}\mathcal{N}(\mathbf{x}|\mathbf{m}, \boldsymbol{\Lambda})$$

no, the mixture of gaussian distribution does not belong to the exponetial family because it cannot be written in the form $p(x|\theta) = \frac{1}{z(\theta)} h(x) exp(dot(t(x), \theta))$. Also, it is a rule that I found on page 6 of this reference:
https://arxiv.org/pdf/0911.4863.pdf

11. [**Bonus**][20 points] Given a distribution in the exponential family $p(\mathbf{x}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ are the natural parameters. As we discussed in the class, the distributions in the exponential family are often parameterized by their expectations, namely $\boldsymbol{\theta} = \mathbb{E}\big(\mathbf{u}(\mathbf{x})\big)$ where $\mathbf{u}(\mathbf{x})$ are the sufficient statistics (recall Gaussian and Bernoulli distributions). Given an arbitrary distribution $p(\mathbf{x}|\boldsymbol{\alpha})$, the Fisher information matrix in terms of the distribution parameters $\boldsymbol{\alpha}$ is defined as $\mathbf{F}(\boldsymbol{\alpha}) = \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\alpha})}[-\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\alpha}))}{\partial \boldsymbol{\alpha}^2}]$.

   (a) [5 points] Show that if we calculate the Fisher Information matrix in terms of the natural parameters, we have $\mathbf{F}(\boldsymbol{\eta}) = \text{cov}\big(\mathbf{u}(\mathbf{x})\big)$.

   (b) [5 points] Show that $\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} = \mathbf{F}(\boldsymbol{\eta})$.

   (c) [10 points] Show that the Fisher information matrix in terms of the expectation parameters is the inverse of that in terms of the natural parameters, $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{F}^{-1}(\boldsymbol{\eta})$.

(d) [5 points] Suppose we observed dataset . Show that

$$\frac{\partial \log p(|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \mathbf{F}(\boldsymbol{\eta})^{-1} = \frac{\partial \log p(|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

and

$$\frac{\partial \log p(|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{F}(\boldsymbol{\theta})^{-1} = \frac{\partial \log p(|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}.$$

Note that I choose the orientation of the gradient vector to be consistent with Jacobian. So, in this case, the gradient vector is a row vector (rather than a column vector). If you want to use a column vector to represent the gradient, you can move the information matrix to the left. It does not influence the conclusion.

1 Practice [20 points]

Code for the practice section can be found at https://github.com/jakehirst/Prob-ML

1. [5 Points] Look into the student t's distribution. Let us set the mean and precision to be $\mu = 0$ and $\lambda = 1$. Vary the degree of freedom $\nu = 0.1, 1, 10, 100, 10^6$ and draw the density of the student t's distribution. Also, draw the density of the standard Gaussian distribution $\mathcal{N}(0, 1)$. Please place all the density curves in one figure. Show the legend. What can you observe?

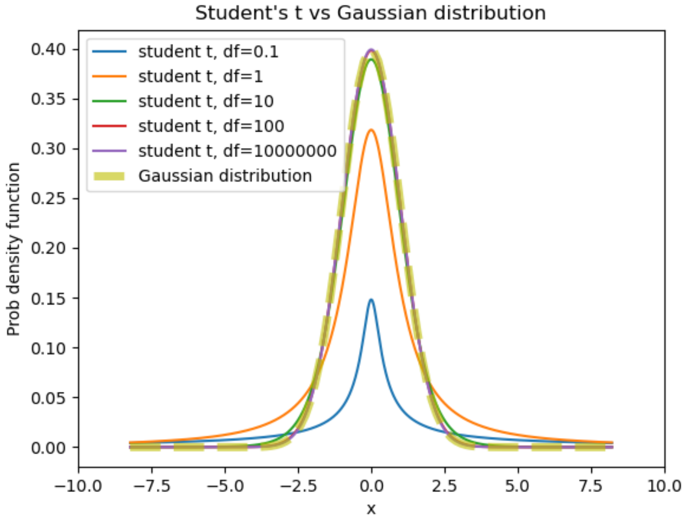

Figure 1: plot for student's t vs Gaussian distributions

As you can see from the plot, it seems as though as the degrees of freedom increase for the student's t distribution, the probability density function looks more and more like a Gaussian distribution.

2. [5 points] Draw the density plots for Beta distributions: Beta(1,1), Beta(5, 5) and Beta (10, 10). Put the three density curves in one figure. What do you observe? Next draw the density plots for Beta(1, 2), Beta(5,6) and Beta(10, 11). Put the three density curves in another figure. What do you observe?

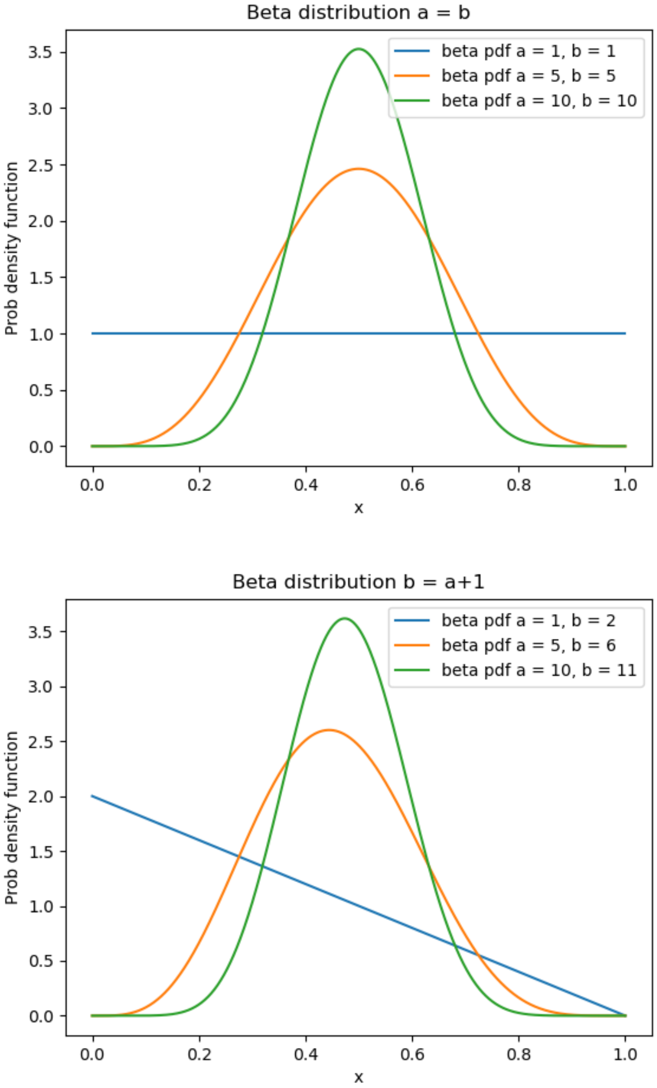

Figure 2: plots for Beta distribution

In both cases, as a increases, the pdf becomes much closer to a bell shaped curve, almost like a normal distribution. In the case where a = b, the curves are always centered around x = 0.5, whereas in the case where b = a+1, the curves are a bit scewed to the left.

3. [10 points] Randomly draw 30 samples from a Gaussian distribution $\mathcal{N}(0, 2)$. Use the 30 samples as your observations to find the maximum likelihood estimation (MLE) for a Gaussian distribution and a student t distribution. For both distributions, please use L-BFGS to optimize the parameters. For student t, you need to estimate the degree of the freedom ν as well. Draw a plot of the estimated the Gaussian distribution density, student t density and the scatter data points. What do you observe, and why? Next, we inject three noises into the data: we append $\{8, 9, 10\}$ to the 30 samples. Find the MLE for the Gaussian and student t distribution again. Draw the density curves and scatter data points in another figure. What do you observe, and why?

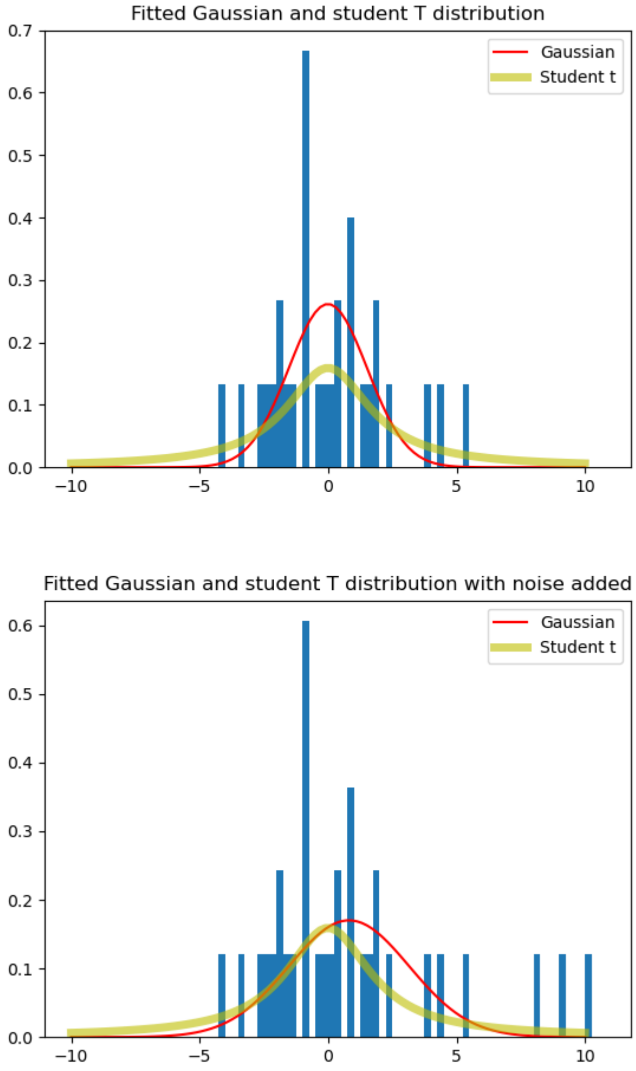

Figure 3: Student t vs Gaussian distributions when noise is added

Since the student t distribution has the heavier tails than a gaussian distribution, it is more resistant to outliers (or noise) in the data. This is shown in the figures above. You can see that when there is no noise added, the student t distribution and the gaussian distribution both capture the data well. However, when noise is added, the gaussian distribution gets skewed towards the noise, while the student t distribution stays relatively the same, still capturing the important distribution of the data.