Reading: Sec 2.8, or some related Wikipedia pages if needed.

1. (Entropy)
   (a) Given $\mathbf{x} \in \mathbb{R}^p$, with density or point mass function $p(\mathbf{x})$, write down the general definition of *entropy*.
   (b) Write down explicit formula for discrete case and continuous case. [Hint: What is the expression for expectation under these two cases?]
   (c) Give interpretation to entropy. (What does it measure/represent?)
   (d) Try to compute the entropy for $\mathbf{x}$ following these distributions:
   - Uniform on a set of size m
   - Bernoulli
   - Geometric
   - univariate normal $N(\mu, \sigma^2)$
   - multivariate normal $N(\boldsymbol{\mu}, \Sigma)$

2. (KL divergence)
   (a) What is a convex function? What is the Jensen's inequality?
   (b) Let P and Q be two distributions defined on the same sample space $\mathcal{S}$, with corresponding density functions or point mass functions being $p(\mathbf{x})$ and $q(\mathbf{x})$ respectively. Write down the definition of *KL-Divergence* between the two distribution.
   (c) Write down explicit formula for discrete case and continuous case. [Hint: What is the expression for expectation under these two cases?]
   (d) Give interpretation to KL divergence. (What does it measure/represent?)
   (e) Show the following properties using the definition:
   - If $P = Q$, then $D_{\mathrm{kl}}(P\|Q) = 0$
   - If $P \neq Q$, then $D_{\mathrm{kl}}(P\|Q) \geq 0$ [Hint: Use Jensen's inequality]

   (f) Use KL divergence to show the following statements:
   - The uniform distribution has the highest entropy over all distributions on the set S.
   - The Gaussian random variables have the largest entropy.

3. (Maximum Likelihood Estimation)
   (a) What is a likelihood function?
   (b) What does MLE do? What's the idea? Describe the estimation procedure using MLE.
   (c) Review some examples on Wikipedia `https://en.wikipedia.org/wiki/Maximum_likelihood_estimation#Examples` if needed.
   (d) Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is from distribution P with density $p(\mathbf{x})$.
   Show that if we adopt the distribution family Q_θ with density function $q_\theta(\mathbf{x})$ to calculate the MLE, then the MLE tries to find the distribution Q_θ that is closest to the true distribution P in the sense of KL divergence.