Tutorial 5 Exercises

1. **Bias, Variance, and Bayes Error.** The purpose of this exercise is to show a simple example where you can compute the bias, variance, and Bayes error of a predictor. For this question, we assume we have N scalar-valued observations $\{x^{(i)}\}_{i=1}^N$ sampled independently from a Gaussian distribution $\mathcal{N}(x; \mu, \sigma^2)$ with known variance σ^2 and unknown mean μ. We'd like to estimate the mean parameter μ, or equivalently, choose a $\hat{\mu}$ which minimizes the squared error risk $\mathbb{E}[(x - \hat{\mu})^2]$.

   We'll introduce the Gaussian distribution properly in a later lecture, but hopefully you've seen it before in a probability course. It is a bell-shaped distribution whose density is:

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

   The details of the Gaussian distribution (such as the density) aren't important for this exercise. The important facts are that $\mathbb{E}[x] = \mu$ and $\mathrm{Var}(x) = \sigma^2$).

   We will estimate the unknown mean paramter μ by taking the empirical mean, or average, of the observations:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}.$$

   This is equivalent to the maximum likelihood estimate, but you don't need to know that yet. (It's covered in a later lecture.)

   The squared error risk $\mathbb{E}[(x - \hat{\mu})^2]$ of this estimator can be decomposed into terms for bias, variance, and Bayes error, exactly following our proof from Lecture 5. (Here, x plays the role of $\mathbf{t}$, and $\hat{\mu}$ plays the role of $\mathbf{y}$. The Bayes optimal prediction, corresponding to $y_\star$ from lecture, is $\mathbb{E}[x] = \mu$.) Your job is to determine each of the three terms.

   (a) Bayes error: $\mathbb{E}[(x - \mu)^2]$

   (b) Bias: $(\mathbb{E}[\hat{\mu}] - \mu)^2$

   (c) Variance: $\mathrm{Var}(\hat{\mu})$

2. **Information Theory.** The goal of this question is to help you become more familiar with the basic equalities and inequalities of information theory. They appear in many contexts in machine learning and elsewhere, so having some experience with them is quite helpful. We review some concepts from information theory, and ask you a few questions.

   Recall the definition of the entropy of a discrete random variable X with probability mass function p:

   $$H(X) = \sum_x p(x) \log_2 \left(\frac{1}{p(x)} \right).$$

   Here the summation is over all possible values of $x \in \mathcal{X}$, which (for simplicity) we assume is finite. For example, $\mathcal{X}$ might be $\{1, 2, \ldots, N\}$. Recall also the definition of conditional entropy:

   $$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x).$$

   (a) Prove that the entropy $H(X)$ is non-negative.

   (b) Prove the Chain Rule for entropy:

   $$H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$$

   (c) Prove that $H(X, Y) \geq H(X)$. *(Hint: this follows fairly directly from parts (a) and (b).)*