CSC311H1F Tutorial 5

Exercises on Bias-Variance Decomposition and Entropy

Ian Shi & Skylar Hao

Overview

- Recap: Generalization error can be decomposed into bias, variance and Bayes error terms.
- Q1: Decompose a predictor for the sample mean estimator of a Gaussian distribution
- Q2: Prove some properties of Entropy

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:

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

Q1: Decomposition

- Decompose the mean squared error (MSE) of sample mean.

$$\mathbb{E}[(x - \hat{\mu})^2]$$

- Take expectation w.r.t. $x \sim N(x; \mu, \sigma^2)$

$$\begin{aligned}
\mathbb{E}_x[(x - \hat{\mu})^2] &= \mathbb{E}[x^2 - 2x\hat{\mu} + \hat{\mu}^2] \\
&= \mathbb{E}[x^2] - 2\hat{\mu}\mathbb{E}[x] + \hat{\mu}^2 \\
&= \mathrm{Var}[x] + \mathbb{E}[x]^2 - 2\hat{\mu}\mathbb{E}[x] + \hat{\mu}^2 \\
&= (\mathbb{E}[x] - \hat{\mu})^2 + \mathrm{Var}[x] \\
&= (\mu - \hat{\mu})^2 + \mathrm{Var}[x]
\end{aligned}$$

Q1: Decomposition

- Take expectation w.r.t estimator $\hat{\mu}$
  - Estimator is a random variable since the training data its generated from is randomly drawn from the true distribution

$$
\begin{aligned}
\mathbb{E}_{\hat{\mu}}[\mathbb{E}_x[(x-\hat{\mu})^2]] &= \mathbb{E}\big[(\mu-\hat{\mu})^2 + \mathrm{Var}[x]\big] \\
&= \mathbb{E}[(\mu-\hat{\mu})^2] + \mathrm{Var}[x] \\
&= \mathbb{E}[(\mu^2 - 2\mu\hat{\mu} + \ \hat{\mu}^2)] + \mathrm{Var}[x] \\
&= \mu^2 - 2\mu\mathbb{E}[\hat{\mu}] + \mathbb{E}[\hat{\mu}^2] + \mathrm{Var}[x] \\
&= \mu^2 - 2\mu\mathbb{E}[\hat{\mu}] + \mathbb{E}[\hat{\mu}]^2 + \mathrm{Var}[\hat{\mu}] + \mathrm{Var}[x] \\
&= (\mu - \mathbb{E}[\hat{\mu}])^2 + \mathrm{Var}[\hat{\mu}] + \mathrm{Var}[x]
\end{aligned}
$$

Q1: Problem Statement

- Find exact bias, variance, Bayes error of sample mean MSE
  - Bias: $(\mu - \mathbb{E}[\hat{\mu}])^2$
  - Variance: $\mathrm{Var}[\hat{\mu}]$
  - Bayes Error: $\mathbb{E}(x - \mu)^2$

- Use properties of expectation / variance
- Remember that $\mathbb{E}[x] = \mu, \mathrm{Var}[x] = \sigma^2$
- Also remember $\hat{\mu}$ is our sample mean estimator, meaning its defined by the equation in the handout

Q1: Bias Solution

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

Looks like we need $\mathbb{E}[\hat{\mu}]$

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[x_i] = \frac{1}{N}\sum_{i=1}^{N} \mu = \frac{1}{N}(N\mu) = \mu$$

Substituting back in

$$(\mu - \mathbb{E}[\hat{\mu}])^2 = (\mu - \mu)^2 \boxed{= 0}$$

Q1: Bias Solution

- Since $(\mu - \mathbb{E}[\hat{\mu}])^2 \; = 0$, it is an <u>unbiased</u> estimator

- Estimators which have bias $= 0$ are unbiased, and vice versa
  - Example of biased estimator: Trying to estimate an unknown variance via

$$S^2 = \frac{1}{n}\sum_i (x_i - \hat{\mu})^2$$

Q1: Variance Solution

$$\mathrm{Var}[\hat{\mu}] = \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N^2}\mathrm{Var}\left[\sum_{i=1}^{N} x_i\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}[x_i]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 = \frac{1}{N^2}(\mathrm{N}\,\sigma^2)$$

- Aside: This can be converted into the standard error formula by square rooting both sides. Pretty cool connection!

Q1: Bayes Error Solution

- Note that we already obtained Bayes error of $\mathrm{Var}[x] = \sigma^2$ in decomposition. Starting from handout equation…

$$\begin{aligned}
\mathbb{E}(x-\mu)^2 &= \mathbb{E}[x^2 - 2x\mu + \mu^2] \\
&= \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mathbb{E}[\mu^2] \\
&= \mathbb{E}[x]^2 + \mathrm{Var}[x] - 2\mu\mathbb{E}[x] + \mathbb{E}[\mu^2] \\
&= \mu^2 + \sigma^2 - 2\mu\mu + \mu^2 \\
&= 2\mu^2 - 2\mu^2 + \sigma^2 \\
&\boxed{= \sigma^2}
\end{aligned}$$

Q2: Entropy Properties Part (a)

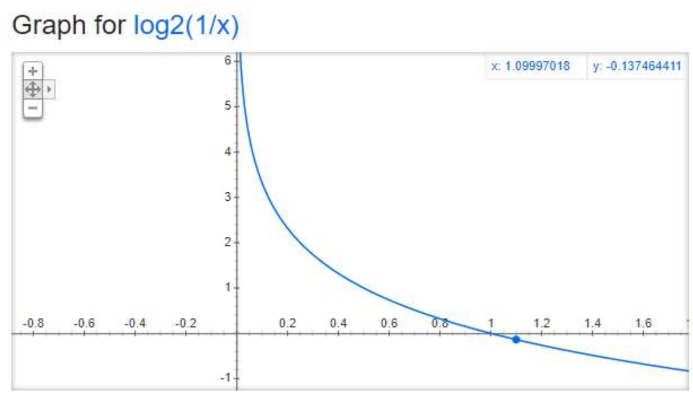
- Prove entropy $H(X)$ is non-negative

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

- X is a discrete random variable. Thus:
  - $p(x_i) \geq 0$
  - $\sum_{x \in \mathcal{X}} p(x) = 1$
- The two conditions also imply $p(x_i) \leq 1$

Q2: Entropy Properties Part (a)

- Since $0 \leq p(x_i) \leq 1, \log_2\left(\frac{1}{p(x)}\right) \geq 0$



- We are basically done.
- $H(X) = \sum_x p(x) \log_2\left(\frac{1}{p(x)}\right)$

Non-negative Non-negative

Sums of non-negative values will remain non-negative

Q2: Entropy Properties Part (b)

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

$$H(X,Y) = \sum_x \sum_y p(x,y)\log_2\left(\frac{1}{p(x,y)}\right)$$

$$= -\sum_x \sum_y p(x,y)\log_2 p(x,y)$$

$$= -\sum_x \sum_y p(\,x,y)\,\log(p(y|x)p(x))$$

$$= -\sum_x \sum_y p(x,y)(\log p(y|x) + \log p(x))$$

Log product identity

$$= -\sum_x \sum_y p(x,y)\log p(y|x)\; - \sum_x \sum_y p(x,y)\log p(x)$$

By commutativity and associativity of summation

Q2: Entropy Properties Part (b)

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(y|x) \; - \sum_x \sum_y p(x,y) \log p(x)$$

$$= -\sum_x \sum_y p(x,y) \log p(y|x) - \sum_x \log p(x) \sum_y p(x,y)$$

Since $\log p(x)$ is not dependent on y

$$= -\sum_x \sum_y p(x,y) \log p(y|x) - \sum_x \log p(x) \left(p(x)\right)$$

Marginalizing out y

$$= -\sum_x \sum_y p(x,y) \log p(y|x) + H(X)$$

By definition of $H(X)$

Q2: Entropy Properties Part (b)

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(y|x) + H(X)$$

$$= -\sum_x \sum_y p(y|x) p(x) \log p(y|x) + H(X)$$

$$= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) + H(X)$$

Since $p(x)$is not dependent on y

$$= -\sum_x p(x) \left(-H(Y|X=x)\right) + H(X)$$

By definition of $H(Y|X=x)$

To show the other way around, we can do equivalent proof, but note $H(Y|X) \neq H(X|Y)$ in general.

Q2: Entropy Properties Part (c)

- Prove $H(X,Y) \geq H(X)$
- We know that $H(X) \geq 0$, and $H(X,Y) = H(Y|X) + H(X)$
- Non rigorous demonstration
  - If $H(Y|X) = 0$, then $H(X,Y) = H(X)$
  - If $H(Y|X) > 0$, then $H(X,Y) \geq H(X,Y) - H(Y|X) = H(X)$
  - $H(Y|X)$ cannot be less than 0 [proof similar to part (a)]