

# Information Theory and Linear Regression

---

CSC311 Summer 2026

University of Toronto

# Information Theory

---

How do we choose between splits when constructing decision trees?

- Measure how much information we can gain from a given split.
- This quantity is call *Information Gain*!
- It is an information theoretic concept that quantifies for a r.v. how much uncertainty is removed if we know its value.

Let's review some information theory basics and definitions.

# Uncertainty and Entropy

Uncertainty is the main building block of many information theory concepts.

- We don't always have all the information about all the variables we care about.
- We use probabilities about events to make *informed* guesses.
- As we learn more information, we can increase confidence, or decrease uncertainty, in our guess.

# Uncertainty and Entropy

- Uncertainty is the main building block of many information theory concepts.
- This uncertainty is quantified as Entropy of the random variable,  $H(X)$ . Mathematically,

For a discrete r.v.:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

For a continuous r.v.:

$$H(X) = - \int_{\mathcal{X}} p(x) \log_2 p(x) dx$$

# Joint Entropy

- We might be interested in the uncertainty in two or more r.v.s that have some joint distribution.
- This is quantified as the Joint Entropy of the r.v.s in question.
- Its mathematical definition follows analogously to that of entropy but with joint probabilities.

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y)$$

Exercise: Can you write down the continuous version of this definition?

# Conditional Entropy

- We are often interested in the uncertainty in one r.v. once we know the value of another.
- This is quantified as the Conditional Entropy of the first *given* the second.
- Its mathematical definition follows analogously to that of entropy with conditional probabilities.

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

# Conditional Entropy

We can expand the terms further:

$$\begin{aligned} H(Y|X) &= - \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) p(y|x) \log_2 p(y|x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(y|x) \end{aligned}$$

Exercise: Continuous version?

## Aside: Logarithm Properties

Some useful properties of logs

- $\log(ab) = \log a + \log b$
- $\log(a/b) = \log a - \log b$

For instance, in the previous slide we encountered  $\log_2 p(y|x)$  which can be written as

$$\log_2 \frac{p(x, y)}{p(x)} = \log_2 p(x, y) - \log_2 p(x)$$

## Information Gain

Finally, we can now quantify a notion of Information Gain, aka Mutual Information between r.v.s  $X$  and  $Y$ .

- This quantifies how much more certain (or less uncertain) we are about  $Y$  if we know the value of  $X$ .
- In other words, how much uncertainty (or entropy) is reduced in  $Y$  once we are *given*  $X$ ?
- Definition: take the entropy of  $Y$  and subtract the conditional entropy of  $Y$  given  $X$ .

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

## Exercises: Information Theory

We now practice computing some of these quantities and prove some standard equalities and inequalities of information theory, which appear in many contexts in machine learning and elsewhere.

## Exercise 1

Let  $p(x, y)$  be given by

	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

Compute

- $H(X), H(Y)$
- $H(X|Y), H(Y|X)$
- $H(X, Y)$
- $IG(Y|X)$

## Exercise 1

Let  $p(x, y)$  be given by

	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

Compute

- $H(X), H(Y)$
- $H(X|Y), H(Y|X)$
- $H(X, Y)$
- $IG(Y|X)$

**(Solution:**  $H(X) = H(Y) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3 = 0.918$

$$H(X|Y) = \frac{1}{3}H(X|Y=0) + \frac{2}{3}H(X|Y=1) = 0.667 = H(Y|X)$$

$$H(X, Y) = 3 \cdot \frac{1}{3}\log 3 = 1.585$$

$$IG(Y|X) = H(Y) - H(Y|X) = 0.251$$

)

## Exercise 2

Prove that entropy  $H(X)$  is non-negative, i.e.,  $H(X) \geq 0$ .  
For reference, we can use the discrete definition:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

## Exercise 2

Prove that entropy  $H(X)$  is non-negative, i.e.,  $H(X) \geq 0$ .  
For reference, we can use the discrete definition:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

(Solution:  $\forall x, p(x) \leq 1 \implies \log_2(p(x)) \leq 0 \implies -p(x) \log_2 p(x) \geq 0$   
 $\implies H(X) \geq 0$   
)

## Exercise 3

Prove the Chain Rule for entropy, i.e.

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

## Exercise 3

Prove the Chain Rule for entropy, i.e.

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

(Solution:  $H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y)$   
 $= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x|y)p(y)$   
 $= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x|y) - \sum_{y \in \mathcal{Y}} \log_2 p(y) \sum_{x \in \mathcal{X}} p(x, y)$   
 $= H(X|Y) - \sum_{y \in \mathcal{Y}} p(y) \log_2 p(y)$   
 $= H(X|Y) + H(Y)$

(Similarly for the other direction.)

)

## Exercise 4

Prove that  $H(X, Y) \geq H(X)$ .

*Hint: you can use results of the first two exercises.*

## Exercise 4

Prove that  $H(X, Y) \geq H(X)$ .

*Hint: you can use results of the first two exercises.*

**(Solution:  $H(X, Y) = H(Y|X) + H(X) \geq H(X)$  since  $H(Y|X) \geq 0$ .  
)**

# Linear Regression

---

# Linear Regression Review

Linear Regression is the problem of predicting a target variable  $y$  as a linear combination of input features  $\mathbf{x}$ .

Fixed inputs given to us:

- Features:  $\mathbf{x} = (x_1, x_2, \dots, x_D) \in \mathbb{R}^D$
- Targets:  $t \in \mathbb{R}$

Parameters that we initialize and learn:

- Weights:  $\mathbf{w} = (w_1, w_2, \dots, w_D) \in \mathbb{R}^D$
- Bias:  $b \in \mathbb{R}$

# Data, Parameters and the Model

- Data is provided to us as  $(\mathbf{x}, t)$  tuples.
- Weights and biases,  $\mathbf{w}$  and  $b$ , are parameters we need to learn.
- We model the predictions  $y$  as:

$$\begin{aligned}y = f(\mathbf{x}) &= \sum_{i=1}^D w_i x_i + b \\ &= \mathbf{w}^T \mathbf{x} + b\end{aligned}$$

We need to find  $\mathbf{w}$  and  $b$  such that  $y$  is close to the ground truth  $t$ .

# Objective Function

To learn and evaluate the linear regression model, we need a measure of “closeness”, formally called a Loss or Objective Function, which we need to minimize.

- Squared Error Loss:  $\mathcal{L}(y, t) = \frac{1}{2}(y - t)^2$ .
- For  $N$  data samples, we average the individual losses over all samples:

$$\begin{aligned}\mathcal{J}(\mathbf{w}) &= \frac{1}{2N} \sum_{i=1}^N (y^{(i)} - t^{(i)})^2 \\ &= \frac{1}{2N} \sum_{i=1}^N (\mathbf{w}^T \mathbf{x}^{(i)} + b - t^{(i)})^2\end{aligned}$$

## Exercise: Linear Regression Bias-Variance

Assume the optimal weights are given by  $\mathbf{w}^*$  and for all data samples

$$t^{(i)} = \mathbf{w}^{*T} \mathbf{x}^{(i)} + \epsilon^{(i)}$$

where  $\epsilon^{(i)}$  are independent random noise variables.  
Further, recall that the loss function is given by

$$\mathcal{J}(w) = \frac{1}{2N} \|\mathbf{y} - \mathbf{t}\|^2$$

## Exercise: Linear Regression Bias-Variance

Using the above, derive the bias-variance decomposition for the linear regression problem.

## Exercise: Linear Regression Bias-Variance

Using the above, derive the bias-variance decomposition for the linear regression problem.

(Solution: Similar to previous bias-variance exercise, with updated notation. )