DSC 40A: Theoretical Foundations of Data Science

Chapter 1: Foundational Concepts

June 30, 2025

Chapter Overview

This chapter will cover several foundational concepts and themes which will permeate the rest of our journey.

Chapter Overview

This chapter will cover several foundational concepts and themes which will permeate the rest of our journey.

- **The Modeling method** - a guiding beacon for how data scientists approach their work.

Chapter Overview

This chapter will cover several foundational concepts and themes which will permeate the rest of our journey.

- **The Modeling method** - a guiding beacon for how data scientists approach their work.
- **The Constant Model** - the simplest model of all,

Chapter Overview

This chapter will cover several foundational concepts and themes which will permeate the rest of our journey.

- **The Modeling method** - a guiding beacon for how data scientists approach their work.
- **The Constant Model** - the simplest model of all,
- Different choices of *loss functions*.

Fundamental Ingredients

Every scenario we encounter as data scientists consists of a few fundamental ingredients.

Fundamental Ingredients

Every scenario we encounter as data scientists consists of a few fundamental ingredients.

Example.

- Annabeth is a biologist working at a company that studies infectious diseases in mice.

Fundamental Ingredients

Every scenario we encounter as data scientists consists of a few fundamental ingredients.

Example.

- Annabeth is a biologist working at a company that studies infectious diseases in mice.
- She wants to understand how the mice respond to different drug dosages (measured in *mg per g body weight*) and measures a mouse's response in *hours survived after infection*.

Fundamental Ingredients

Every scenario we encounter as data scientists consists of a few fundamental ingredients.

Example.

- Annabeth is a biologist working at a company that studies infectious diseases in mice.
- She wants to understand how the mice respond to different drug dosages (measured in *mg per g body weight*) and measures a mouse's response in *hours survived after infection*.
- She would like to build a model that, given a specific dosage, predicts the expected number of bacteria in the mouse's blood.

Step 1 - Identify Variables

Input variables (features). The data or attributes which are used to make predictions. They represent the *independent* variables.

Step 1 - Identify Variables

Input variables (features). The data or attributes which are used to make predictions. They represent the *independent* variables.
Output variables (targets). The data or attributes that the model aims to predict - the *dependent* variables.

Step 1 - Identify Variables

Input variables (features). The data or attributes which are used to make predictions. They represent the *independent* variables.
Output variables (targets). The data or attributes that the model aims to predict - the *dependent* variables.

In our example. Dosage $\rightarrow$ input variable.
Survival time $\rightarrow$ output variable.

Step 1 - Identify Variables

Input variables (features). The data or attributes which are used to make predictions. They represent the *independent* variables.
Output variables (targets). The data or attributes that the model aims to predict - the *dependent* variables.

In our example. Dosage $\rightarrow$ input variable.
Survival time $\rightarrow$ output variable.

Variable types.

- **Numerical variables**: height, dosage, weight of an acorn, ...

Step 1 - Identify Variables

Input variables (features). The data or attributes which are used to make predictions. They represent the *independent* variables.
Output variables (targets). The data or attributes that the model aims to predict - the *dependent* variables.

In our example. Dosage $\rightarrow$ input variable.
Survival time $\rightarrow$ output variable.

Variable types.

- **Numerical variables**: height, dosage, weight of an acorn, ...
- **Categorical variables**: cat color, movie genre, college major, ...

Step 1 - Identify Variables

Input variables (features). The data or attributes which are used to make predictions. They represent the *independent* variables.
Output variables (targets). The data or attributes that the model aims to predict - the *dependent* variables.

In our example. Dosage $\rightarrow$ input variable.
Survival time $\rightarrow$ output variable.

Variable types.

- **Numerical variables**: height, dosage, weight of an acorn, ...
- **Categorical variables**: cat color, movie genre, college major, ...
- **Binary variables**: true/false, diseased/healthy, success/failure.

Step 1 - Identify Variables

Input variables (features). The data or attributes which are used to make predictions. They represent the *independent* variables.
Output variables (targets). The data or attributes that the model aims to predict - the *dependent* variables.

In our example. Dosage $\rightarrow$ input variable.
Survival time $\rightarrow$ output variable.

Variable types.

- **Numerical variables**: height, dosage, weight of an acorn, ...
- **Categorical variables**: cat color, movie genre, college major, ...
- **Binary variables**: true/false, diseased/healthy, success/failure.

Concept Check

1. In Annabeth's study, suppose she also records the *infection type* (viral vs. bacterial).
   (a) What **variable type** is this?
   (b) Is it an *input* or an *output* in our current setup?

Concept Check

1. In Annabeth's study, suppose she also records the *infection type* (viral vs. bacterial).
   (a) What **variable type** is this?
   (b) Is it an *input* or an *output* in our current setup?
2. What is another scenario where the input features could consist of multiple types?

Concept Check

1. In Annabeth's study, suppose she also records the *infection type* (viral vs. bacterial).
   (a) What **variable type** is this?
   (b) Is it an *input* or an *output* in our current setup?
2. What is another scenario where the input features could consist of multiple types?
3. What about a scenario where the *output* features consist of multiple types?

Step 2 - Choose a Model

A **model** is a function that maps features to targets based on a set of parameters, aiming to capture the underlying relationship between them.

Step 2 - Choose a Model

A **model** is a function that maps features to targets based on a set of parameters, aiming to capture the underlying relationship between them.

Parameters (or *weights*) are the numerical values that define the model; you can think of them as dials and knobs.

Step 2 - Choose a Model

A **model** is a function that maps features to targets based on a set of parameters, aiming to capture the underlying relationship between them.

Parameters (or *weights*) are the numerical values that define the model; you can think of them as dials and knobs.

Model training is the process of tuning the weights to improve overall accuracy.

Back to our example

Recall: Annabeth is modeling mouse survival times in terms of dosage of a drug.

Back to our example

Recall: Annabeth is modeling mouse survival times in terms of dosage of a drug.

Let x denote the drug dosage (in mg/g). We choose to model the survival time by

$$f(x) = c\,x,$$

for some $c \in \mathbb{R}$.

Back to our example

Recall: Annabeth is modeling mouse survival times in terms of dosage of a drug.

Let x denote the drug dosage (in mg/g). We choose to model the survival time by

$$f(x) = c\,x,$$

for some $c \in \mathbb{R}$. This is a **simple linear model** with a single parameter c.

Back to our example

Recall: Annabeth is modeling mouse survival times in terms of dosage of a drug.

Let x denote the drug dosage (in mg/g). We choose to model the survival time by

$$f(x) = c\,x,$$

for some $c \in \mathbb{R}$. This is a **simple linear model** with a single parameter c.

What are some examples of different models Annabeth could use?

Step 3 - Select a Loss Function

The million dollar question: How should Annabeth choose c to yield accurate predictions?

Step 3 - Select a Loss Function

The million dollar question: How should Annabeth choose c to yield accurate predictions?

She requires a **loss function**: a function that quantifies the difference, or error, between the predicted and actual target values.

Step 3 - Select a Loss Function

The million dollar question: How should Annabeth choose c to yield accurate predictions?

She requires a **loss function**: a function that quantifies the difference, or error, between the predicted and actual target values.

Loss functions assign a numerical value to prediction errors and guide the adjustment of weights. Common choices include **mean-squared error**, **absolute loss**, **cross-entropy loss**, and more.

Square Loss Example

To keep things simple, Annabeth chooses the **square loss**.

Square Loss Example

To keep things simple, Annabeth chooses the **square loss**.

Let c be the parameter, x the mouse's dosage, and y its *actual* survival time. The square loss is given by

$$L\big(c; (x, y)\big) \;=\; (\,y - c\,x\,)^2.$$

Square Loss Example

To keep things simple, Annabeth chooses the **square loss**.

Let c be the parameter, x the mouse's dosage, and y its *actual* survival time. The square loss is given by

$$L\big(c;\,(x,y)\big) \;=\; (\,y - c\,x\,)^2.$$

Here $L(c;\,(x,y))$ means **"the loss associated with parameter** c **on the example** (x,y)**."** The semicolon separates the parameter from the input-output pair.

Square Loss Example

To keep things simple, Annabeth chooses the **square loss**.

Let c be the parameter, x the mouse's dosage, and y its *actual* survival time. The square loss is given by

$$L(c; (x, y)) \;=\; (\, y - c\, x\,)^2.$$

Here $L(c; (x, y))$ means **"the loss associated with parameter c on the example (x, y)."** The semicolon separates the parameter from the input-output pair.
When $y \approx c\, x$ the loss is small; otherwise it grows quadratically with their difference.

Square Loss Example

To keep things simple, Annabeth chooses the **square loss**.

Let c be the parameter, x the mouse's dosage, and y its *actual* survival time. The square loss is given by

$$L\big(c;\,(x,y)\big) \;=\; \big(\,y - c\,x\,\big)^2.$$

Here $L(c;\,(x,y))$ means **"the loss associated with parameter** c **on the example** (x,y)**."** The semicolon separates the parameter from the input-output pair.

When $y \approx c\,x$ the loss is small; otherwise it grows quadratically with their difference.

Choosing an appropriate loss function depends on the problem and desired behavior such as robustness to outliers (more on this to come...).

Minimizing the Loss: One Mouse

If Annabeth has a *single* mouse with variables (x_1, y_1), she can minimize

$$L(c;\,(x_1, y_1)) \;=\; \big(y_1 - c\,x_1\big)^2$$

with respect to c, as follows...

Minimizing the Loss: One Mouse

If Annabeth has a *single* mouse with variables (x_1, y_1), she can minimize

$$L(c;\, (x_1, y_1)) \;=\; \big(y_1 - c\, x_1\big)^2$$

with respect to c, as follows...
Differentiate with respect to c:

$$\begin{aligned}\frac{\mathrm{d}L}{\mathrm{d}c} &= \frac{\mathrm{d}}{\mathrm{d}c}\big(y_1 - cx_1\big)^2 \\ &= 2\big(y_1 - cx_1\big)\underbrace{\frac{\mathrm{d}}{\mathrm{d}c}(y_1 - cx_1)}_{=-x_1} \\ &= -2x_1\big(y_1 - cx_1\big).\end{aligned}$$

Critical Point and Optimal Value

Set the derivative to zero to locate a critical point c^*:

Critical Point and Optimal Value

Set the derivative to zero to locate a critical point c^*:

$$-2x_1\left(y_1 - c^* x_1\right) = 0 \implies y_1 - c^* x_1 = 0 \implies c^* = \frac{y_1}{x_1}.$$

Critical Point and Optimal Value

Set the derivative to zero to locate a critical point c^*:

$$-2x_1\big(y_1 - c^* x_1\big) = 0 \implies y_1 - c^* x_1 = 0 \implies c^* = \frac{y_1}{x_1}.$$

Taking a look at the second derivative:

$$\frac{\mathrm{d}^2 L}{\mathrm{d}c^2} = \frac{\mathrm{d}}{\mathrm{d}c}\big[-2x_1(y_1 - cx_1)\big] = 2x_1^2 \ > \ 0,$$

so c^* indeed *minimizes* the loss.

Critical Point and Optimal Value

To summarize: If Annabeth has a single mouse with dosage x_1 and survival time y_1, the "best" choice of c in the survival time model $f(x) = cx$ is given by $c^* = y_1/x_1$.

Training vs. Testing Data

But what if Annabeth has ten mice instead of one?

Training vs. Testing Data

But what if Annabeth has ten mice instead of one?

- **Training data**: the data used to train parameters and build an accurate model. Here it might look like a collection of points of the form $\{(x_i, y_i)\}_{i=1}^{10}$.

Training vs. Testing Data

But what if Annabeth has ten mice instead of one?

- **Training data**: the data used to train parameters and build an accurate model. Here it might look like a collection of points of the form $\{(x_i, y_i)\}_{i=1}^{10}$.
- **Testing data**: the data which is held back to evaluate performance.

Empirical Risk Function

To use *all* training data we build a **risk function**, which is a function that combines loss values across an entire dataset.

Empirical Risk Function

To use *all* training data we build a **risk function**, which is a function that combines loss values across an entire dataset.

The **empirical risk function** is just the average across the loss values and is given by:

$$\begin{aligned} R(c;\, \{(x_i, y_i)\}_{i=1}^{10}) &= \frac{1}{10} \sum_{i=1}^{10} L\big(c;\, (x_i, y_i)\big) \\ &= \frac{1}{10} \sum_{i=1}^{10} \big(y_i - c x_i\big)^2. \end{aligned}$$

Empirical Risk Function

To use *all* training data we build a **risk function**, which is a function that combines loss values across an entire dataset.

The **empirical risk function** is just the average across the loss values and is given by:

$$R(c;\, \{(x_i, y_i)\}_{i=1}^{10}) = \frac{1}{10}\sum_{i=1}^{10} L\big(c;\, (x_i, y_i)\big)$$
$$= \frac{1}{10}\sum_{i=1}^{10} \big(y_i - cx_i\big)^2.$$

Sometimes we abbreviate this $R(c)$ when the dataset is clear.

Derivative of the Empirical Risk

Compute

$$\frac{\mathrm{d}R}{\mathrm{d}c} = \frac{1}{10}\sum_{i=1}^{10} 2\big(y_i - cx_i\big)(-x_i) = -\frac{2}{10}\sum_{i=1}^{10}\big(x_i y_i - cx_i^2\big).$$

Derivative of the Empirical Risk

Compute

$$\frac{\mathrm{d}R}{\mathrm{d}c} = \frac{1}{10}\sum_{i=1}^{10} 2\big(y_i - cx_i\big)\big(-x_i\big) = -\frac{2}{10}\sum_{i=1}^{10}\big(x_i y_i - cx_i^2\big).$$

Factor:

$$\frac{\mathrm{d}R}{\mathrm{d}c} = -\frac{2}{10}\left(\sum_{i=1}^{10} x_i y_i - c\sum_{i=1}^{10} x_i^2\right).$$

Derivative of the Empirical Risk

Compute

$$\frac{\mathrm{d}R}{\mathrm{d}c} = \frac{1}{10}\sum_{i=1}^{10} 2(y_i - cx_i)(-x_i) = -\frac{2}{10}\sum_{i=1}^{10}(x_i y_i - cx_i^2).$$

Factor:

$$\frac{\mathrm{d}R}{\mathrm{d}c} = -\frac{2}{10}\left(\sum_{i=1}^{10} x_i y_i - c\sum_{i=1}^{10} x_i^2\right).$$

Set $\frac{\mathrm{d}R}{\mathrm{d}c} = 0 \Longrightarrow$

$$c^* = \frac{\displaystyle\sum_{i=1}^{10} x_i y_i}{\displaystyle\sum_{i=1}^{10} x_i^2}.$$

Key Idea: The Modeling Method

1. Identify input and output variables.

Key Idea: The Modeling Method

1. Identify input and output variables.
2. Choose a model.

Key Idea: The Modeling Method

1. Identify input and output variables.
2. Choose a model.
3. Choose a loss and a risk function.

Key Idea: The Modeling Method

1. Identify input and output variables.
2. Choose a model.
3. Choose a loss and a risk function.
4. Find a minimizer of the risk.