

Lecture 13

Feature engineering and transformations

DSC 40A, Fall 2024

Announcements

- Homework 3 is due today.
- Homework 2 scores will be available on Gradescope this weekend.
- Midterm logistics will be announced on Monday.
 - Additional office hours next week for the midterm.
 - Prepare by practicing with old exam problems at practice.dsc40a.com.
 - Problems are sorted by topic!

Agenda

- Feature engineering and transformations.

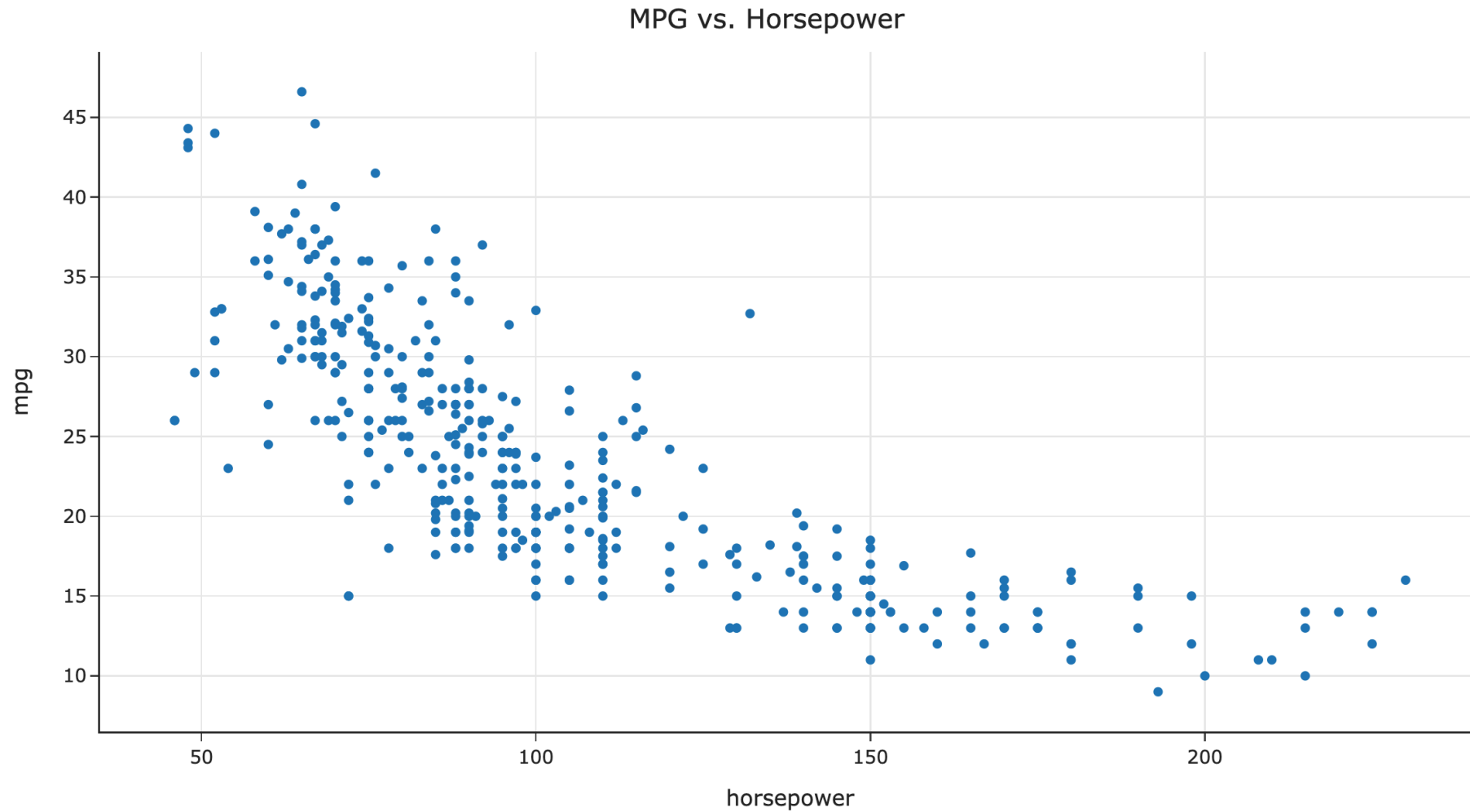
Question 🤔

Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!

If the direct link doesn't work, click the "🤔 Lecture Questions"
link in the top right corner of dsc40a.com.

Feature engineering and transformations



Question: Would a linear hypothesis function work well on this dataset?

A quadratic hypothesis function

- It looks like there's some sort of quadratic relationship between horsepower and MPG in the last scatter plot. We want to try and fit a hypothesis function of the form:

$$H(x) = w_0 + w_1x + w_2x^2$$

- Note that while this is quadratic in horsepower, it is **linear in the parameters!**
 - That is, it is a **linear combination of features**.
- We can do that, by choosing our two "features" to be x_i and x_i^2 , respectively.
 - In other words, $x_i^{(1)} = x_i$ and $x_i^{(2)} = x_i^2$.
 - More generally, we can create new features out of existing features.

A quadratic hypothesis function

- Desired hypothesis function: $H(x) = w_0 + w_1x + w_2x^2$.
- The resulting design matrix looks like:

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & & \\ 1 & x_n & x_n^2 \end{bmatrix}$$

- To find the optimal parameter vector \vec{w}^* , we need to solve the **normal equations!**

$$X^T X \vec{w}^* = X^T \vec{y}$$

More examples

- What if we want to use a hypothesis function of the form:

$$H(x) = w_0 + w_1x + w_2x^2 + w_3x^3?$$

- What if we want to use a hypothesis function of the form:

$$H(x) = w_1 \frac{1}{x^2} + w_2 \sin x + w_3 e^x?$$

Feature engineering

- The process of creating new features out of existing information in our dataset is called **feature engineering**.
- In this class, feature engineering will mostly be restricted to creating non-linear functions of existing features (as in the previous example).
- In the future you'll learn how to do other things, like encode categorical information.
 - You'll be exposed to this in Homework 4, Problem 5!

Non-linear functions of multiple features

- Recall our earlier example of predicting sales from square footage and number of competitors. What if we want a hypothesis function of the form:

$$\begin{aligned} H(\text{sqft}, \text{comp}) &= w_0 + w_1 \cdot \text{sqft} + w_2 \cdot \text{sqft}^2 + w_3 \cdot \text{comp} + w_4 \cdot (\text{sqft} \cdot \text{comp}) \\ &= w_0 + w_1 s + w_2 s^2 + w_3 c + w_4 sc \end{aligned}$$

- The solution is to choose a design matrix accordingly:

$$X = \begin{bmatrix} 1 & s_1 & s_1^2 & c_1 & s_1 c_1 \\ 1 & s_2 & s_2^2 & c_2 & s_2 c_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_n & s_n^2 & c_n & s_n c_n \end{bmatrix}$$

Finding the optimal parameter vector, \vec{w}^*

- As long as the form of the hypothesis function permits us to write $\vec{h} = X\vec{w}$ for some X and \vec{w} , the mean squared error is:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$$

- Regardless of the values of X and \vec{y} , the value of \vec{w}^* that minimizes $R_{\text{sq}}(\vec{w})$ is the solution to the **normal equations**:

$$X^T X \vec{w}^* = X^T \vec{y}$$

Linear in the parameters

- We can fit rules like:

$$w_0 + w_1x + w_2x^2 \quad w_1e^{-x^{(1)2}} + w_2 \cos(x^{(2)} + \pi) + w_3 \frac{\log 2x^{(3)}}{x^{(2)}}$$

- This includes arbitrary polynomials.
- These are all linear combinations of (just) features.
- We can't fit rules like:
$$w_0 + e^{w_1x} \quad w_0 + \sin(w_1x^{(1)} + w_2x^{(2)})$$
 - These are **not** linear combinations of just features!
- We can have any number of parameters, as long as our hypothesis function is **linear in the parameters**, or linear when we think of it as a function of the parameters.

Determining function form

- How do we know what form our hypothesis function should take?
- Sometimes, we know from *theory*, using knowledge about what the variables represent and how they should be related.
- Other times, we make a guess based on the data.
- Generally, start with simpler functions first.
 - Remember, the goal is to find a hypothesis function that will generalize well to unseen data.

Example: Amdahl's Law

- Amdahl's Law relates the runtime of a program on p processors to the time to do the sequential and nonsequential parts on one processor.

$$H(p) = t_S + \frac{t_{NS}}{p}$$

- Collect data by timing a program with varying numbers of processors:

Processors	Time (Hours)
1	8
2	4
4	3

Example: Fitting $H(x) = w_0 + w_1 \cdot \frac{1}{x}$

Processors	Time (Hours)
1	8
2	4
4	3

How do we fit hypothesis functions that aren't linear in the parameters?

- Suppose we want to fit the hypothesis function:

$$H(x) = w_0 e^{w_1 x}$$

- This is **not** linear in terms of w_0 and w_1 , so our results for linear regression don't apply.
- **Possible solution:** Try to apply a **transformation**.

Transformations

- **Question:** Can we re-write $H(x) = w_0 e^{w_1 x}$ as a hypothesis function that is linear in the parameters?

Transformations

- **Solution:** Create a new hypothesis function, $T(x)$, with parameters b_0 and b_1 , where $T(x) = b_0 + b_1x$.
- This hypothesis function is related to $H(x)$ by the relationship $T(x) = \log H(x)$.
- \vec{b} is related to \vec{w} by $b_0 = \log w_0$ and $b_1 = w_1$.
- Our new observation vector, \vec{z} , is
$$\begin{bmatrix} \log y_1 \\ \log y_2 \\ \dots \\ \log y_n \end{bmatrix}.$$
- $T(x) = b_0 + b_1x$ is linear in its parameters, b_0 and b_1 .
- Use the solution to the normal equations to find \vec{b}^* , and the relationship between \vec{b} and \vec{w} to find \vec{w}^* .

Once again, let's try it out! Follow along in [this notebook](#).

Non-linear hypothesis functions in general

- Sometimes, it's just not possible to transform a hypothesis function to be linear in terms of some parameters.
- In those cases, you'd have to resort to other methods of finding the optimal parameters.
 - For example, $H(x) = w_0 \sin(w_1 x)$ **can't** be transformed to be linear.
 - But, there are other methods of minimizing mean squared error:

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n (y_i - w_0 \sin(w_1 x))^2$$

- One method: **gradient descent**, the topic of the next lecture!
- Hypothesis functions that are linear in the parameters are much easier to work with.

Question 🤔

Answer at q.dsc40a.com

Which hypothesis function is **not** linear in the parameters?

- A. $H(\vec{x}) = w_1(x^{(1)}x^{(2)}) + \frac{w_2}{x^{(1)}} \sin(x^{(2)})$
- B. $H(\vec{x}) = 2^{w_1}x^{(1)}$
- C. $H(\vec{x}) = \vec{w} \cdot \text{Aug}(\vec{x})$
- D. $H(\vec{x}) = w_1 \cos(x^{(1)}) + w_2 2^{x^{(2)}} \log x^{(3)}$
- E. More than one of the above.

Roadmap

- This is the end of the content that's in scope for the Midterm Exam.
- Now, we'll introduce **gradient descent**, a technique for minimizing functions that can't be minimized directly using calculus or linear algebra.
- After the Midterm Exam, we'll:
 - Switch gears to **probability**.