More on linear models

Herman Kamper

Feature importance based on weights

We can use the feature weights as a crude way to estimate feature importance.

This assumes that features have the same scale:

- If x_1 is in \$ and x_2 in m^2 then w_1 and w_2 will not be comparable.
- If we scale the features appropriately before training, then we can compare the weights.

But keep in mind that just looking at the weights does not actually tell us the anything about causation. E.g. you might have to features encoding very similar things.

With categorical features (sometimes encoded as one-hot or dummy variables) you cannot do the above. Quoting from ISL:

- "However, the coefficients ... do depend on the choice of dummy variable coding."
- "Rather than rely on the individual coefficients, we can use an F-test ..." This test involves comparing the results of a model with the categorical variable to one without it.

The dummy variable trap

We sometimes need to be careful with using one-hot encodings (dummy variables) to encode categorical features, specifically with linear regression

Linear regression recap

$$\boxed{\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}}$$

- Guaranteed to give the least squares solution, if ...
  - If the features are selected so that there is actually a global minimum.
- If you are not careful, you might screw up the features, e.g. by having features that are multicollinear:
  - When this happens, $\mathbf{X}^\top \mathbf{X}$ won't be invertible.
  - Python will cry, so you will know.

Example: The dummy variable trap

We want to predict salary based on occupation (`student`, `lecturer`, `artist`). We one-hot encode the categorical input:

$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

The resulting model:

$$
\begin{aligned}
f(\mathbf{x}) &= w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \\
&= \begin{cases} w_0 + w_1 & \text{if } \texttt{student} \\ w_0 + w_2 & \text{if } \texttt{lecturer} \\ w_0 + w_3 & \text{if } \texttt{artist} \end{cases}
\end{aligned}
$$

This result is problematic:

- Subtract any constant c from w_0 and add c to w_1, w_2 and w_3: this new model will still give exactly the same predictions.
- So there is not a single optimal $\hat{\mathbf{w}}$.
- If you tried to use the normal equations, you will see that $\mathbf{X}^\top \mathbf{X}$ is not invertible.
- This is because we have multicollinearity between x_1, x_2 and x_3.

How do we not fall into the trap?

1. You could remove the bias term w_0 and then everything would work. But not if you have more than one categorical variable that is one-hot encoded.
2. You could assign one of the categories to $\begin{bmatrix} 0 & 0 \end{bmatrix}^\top$, i.e. you have

$$
\begin{aligned}
\texttt{student} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
\texttt{lecturer} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
\texttt{artist} &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\end{aligned}
$$

   I leave it as an exercise to show for yourself that you will get a unique optimal $\hat{\mathbf{w}}$ in this case.

Exercises

Exercise: Multiple categorical variables

We want to predict a student's project mark based on the supervisor identity and the examiner. We have two potential supervisors (Jackie or Nathie) and two potential examiners (Herman or John).

1. Show that if we use two-dimensional one-hot encodings for both the supervisor and examiner, we also fall into the dummy variable trap.

2. Show that by removing the bias term w_0, in this case where we have two one-hot variables, we would still fall into the dummy variable trap.