✓ 100 XP

Improving classification models

6 minutes

In our exercises, we found that our model could predict avalanches so some degree, but it was still wrong around 40% of the time. This error amount is because our feature – the number of weak layers of snow – isn't the only thing that is responsible for avalanches.

There are two primary ways to improve classification model performance that we'll dive into now: providing additional features and being selective about what enters the model.

Provide additional features

Like linear regression, logistic regression doesn't have to be limited to a single input. It can combine features to make predictions. For example, we might try to predict avalanches based on snow fall and number of hikers disturbing a trail. We can enter both of these features into the same model to calculate a probability of an avalanche.

Internally, logistic regression combines features similarly to linear regression. That is, it treats all features as **independent**, meaning that it assumes that features don't influence one another. For example, our model will assume that the amount of snowfall doesn't change how many people will visit the trail. By default, it also assumes that snowfall increases risk of avalanche by a set amount – regardless as to how many hikers are walking the trail.

The good and bad sides of independent features

Logistic regression can be explicitly told to combine features so that how they work together can be modeled, but won't by default. Making logistic regression different from most other well-known categorization algorithms such as *decision trees* and *neural networks*.

The fact that logistic regression treats features as independent by default is both a strength and a limitation that should be kept in mind. For example, this means it can make clear predictions simply, such as “increasing the number of people increases risk”, which cannot usually be done with other models. It also reduces the chance of overfitting the training data. By contrast, the model can fail to work well if features *actually interact in the real-world*. For example, five hikers crossing a mountain is risky if there's snow, but five people is safe if there's no snow-fall to cause an avalanche. A logistic regression model needs to be told explicitly to look for an **interaction** between snow-fall and number of hikers in this example to pick up this nuance.

Think about your features

The other way to improve models is to give real thought to which features are supplied, and why. Generally, the more features we add to a model, the better the model will work. This is only true, however, if the features we provide are actually relevant and explain something that existing features don't.

Avoiding overtraining

If we supply additional features that aren't particularly useful, the model can *overtrain*. Giving the appearance of working better, but actually working worse in the real world.

For example, imagine if we had daily records of the amount_of_snow, number_of_hikers, temperature, and number_of_birds_spotted. The number of birds spotted is probably not relevant information. Yet, if supplied the model will end up modeling a relationship between avalanches and the number of birds spotted on given days. If birds were spotted more on avalanche days, the model will suggest that birds could be responsible for causing avalanches. We might then set up a systematic bird watching program to predict avalanches, only to find it doesn't work at all.

Avoiding undertraining

Using features naively can also lead to *undertraining* and not make predictions as correctly as possible. For example, the temperature and the number_of_hikers might both be strongly linked to avalanches. Yet, if people only walk on sunny days, the model might find it difficult to differentiate how important hikers are in comparison to temperature. Similarly, we probably will find that our model works better if we supply our number_of_hikers as an exact count of visitors, rather than simply “high” or “low”, so that the model training can find a more exact relationship.

Next unit: Exercise - Improving classification models

Continue >