🧬

Module 5

Evaluation and Dataset Optimization: Overfitting, Underfitting, Variance, and Bias

▼ What is bias?

- The bias of a model is its tendency to "miss" patterns in the data.
- high bias = overgeneralize (i.e., oversimplify) the problem
- For many sources of data there will exist algorithms that fail to accurately capture the relationships between variables--for example, linear regression will fail to adequately describe data generated by *y*=*x*3.
- In that case, much of the model error will be due to the choice of a model that never stood a chance of describing the data well, and we call that error *bias,* and we would say that linear regression **underfits** the data--it doesn't have sufficient power to describe such a complex relationship.

▼ What is variance?

- Variance is the tendency for an algorithm to come up with very different-looking models depending on the exact sample of data the model is based on.
- If instead of linear regression, we used a sixth-order polynomial to try to describe data generated by *y*=*x*3 then the resulting model is going to use some of its capacity to "learn" the locations of the points, and the model will depend more on the specific choice of points offered to it than on the underlying cubic relationship.
- In this case the model **overfits** the data.

▼ What's the difference between high bias and high variance?

- Choosing a model with high bias is to bring wrong assumptions to a problem, and choosing a model with high variance is to put too much emphasis on the specific available data points.
- Try to choose the most appropriate model for a problem--one that minimizes the bias and
  the variance.

▼ Why are certain types of models more prone to overfitting?

- They have a higher degree of flexibility when attempting to approximate a target function.
- These include models that are non-parametric and non-linear.
- Example: decision trees
  - Too many branches tends to lead to overfitting to the data.
  - "Pruning" a decision tree: entails optimizing the tree size and number of branches.
  - Too many branches tends to lead to overfitting to the data.
- Example: neural networks
  - This is known to even occur when the dataset is large.
  - Ways to mitigate overfitting with neural networks would be to reduce number of weights and changing network parameters (weight values) of less-important features as well as introducing noise into the dataset.

▼ What is the variance-bias tradeoff?

The variance-bias tradeoff in machine learning is based around the idea of error that can result from either making your model too complex (high bias, low variance) and specific or too simple and general (low bias, high variance).

▼ What does data augmentation do?

- Data augmentation is one way of expanding a small dataset by manipulating the original datapoints to create additional inputs.

    - Not only will data augmentation expand the amount of training data available, but it also has the benefit of improving the robustness of a model by introducing more variability into the input data.

- ▼ How can data be augmented?
    - There are various ways input data can be augmented and it will likely depend on the use case to determine which make the most sense.
    - Using image recognition as an example (because it is the easiest to visualize), simple tweaks to the photos can be made including rotations, crops, zooms, brightness or contrast adjustments and flips.
    - Other more complex augmentations can include things like using GANs to add weather effects or synthetic objects.
    - For other tasks like NLP, data augmentation is still possible, though arguably harder to implement than on images. NLP augmentation can include things like simile swaps, deletion or addition of words, and translating the original input to another language and back.

- ▼ Loss and cost both represent what?

    the error rate

- ▼ What does loss represent?

    the loss function is specifically referring to the error rate over a single training example

- ▼ What is the loss function?

    The loss function is the difference between the predicted value in a machine learning model and the true value.

- ▼ What does cost represent?

    the cost function is referring to the error over an entire training set

- ▼ What is the cost function?
    - The cost function is the average loss of the entire data set, so the difference is just the scope of the calculation.

- For example, if the prediction of the fifth item in a training set is a 1 and the true value is a 6, the loss function is 5 for that evaluation. If the previous 4 predictions were perfect, the cost function would at this point of the training set would be 1 as a cumulative error of 5 divided by the 5 training items means there would be an average error of 1 across the training set.
- It is more beneficial to optimize the cost function rather than the loss function since the average of the data set is likely to be more representative of the data set as a whole over time as the law of averages states most future events are likely to balance any past deviation from a presumed average.

▼ What happens when an error rate is too low?

an error rate that is too low can mean that your model has overfit the data and will be unable to generalize

▼ What does a machine learning model approximate?

an objective function that maps input variables to an output variable

▼ What is overfitting?

- A model has been overtrained on the initial dataset and fails to generalize to new, unseen samples.
- Overfitting can occur when there are too many parameters or when the data is extremely noisy.
- When the cost of your model is very low, it may be overfitting the data or to the validation set.
- A model that is overfitting the data it was trained on has high **variance**, because it will correctly map inputs it has seen before to its output, but will be thrown off as soon as the input changes,

▼ What is underfitting?

- Some models have trouble stabilizing their error rate, giving the model a high cost.
- This can often be from not having enough parameters, a lack of data, or even data that is too "perfect" and thus not noisy enough to correctly represent the

problem case.

- These models display high bias, meaning they overgeneralize.
- If a model is found to be underfitting, you need to reduce the error rate somehow. This often means increasing the parameters of the model or increasing training time.

▼ What is cross-validation?

One strategy that is very helpful in avoiding overfitting is *cross-validation*, of which one form is known as *k-cross-validation*, a resampling technique that trains and tests a model k number of times on a variety of subsets of the training data to improve estimation of error/accuracy on new input samples.

▼ What is regularization?

it adds a penalty to the error function that keeps coefficients with extreme values at bay

▼ What is the Mean Absolute Error (MAE)?

- The Mean Absolute Error (MAE) is the calculated average of the difference (also known as the residual) between the input values and the output values.
- The MAE is a measure of the accuracy of the model's prediction, but does not tell you much more about whether you have overfit or underfit the data.

▼ What are Mean Square Error (MSE) and Root Mean Square Error (RMSE)?

- Similar to MAE, Mean square error (MSE) calculates the average of the difference between the input values and the output values, only squaring the difference first.
- Root mean square error (RMSE) measures the absolute average magnitude of the error by taking it a step further by averaging all of the results of MSE and then taking the square root of that. RMSE enables large errors to be more readily identified.

▼ What is Mean Absolute Percentage Error (MAPE)?

- Mean Absolute Percentage Error (MAPE), also sometimes called mean absolute percentage deviation (MAPD), is an average of absolute error as a percentage.

- The residual is subtracted from this average.

▼ What are measurements of true and false positives and negatives?

Positive Predictive Value (PPV), Specificity or True Negative Rate (TNR), True Positive Rate (TPR), and False Positive Rate (FPR) are all measurements of true and false positives and negatives

▼ When you find the harmonic mean, what can you identify?

the direction the model is leaning in terms of precision vs recall

▼ What is the F measure good for measuring?

generalization

▼ How can you identify the area under the curve (AUC)?

By plotting the FPR and TPR together, you can identify the area under the curve. This measure is being re-examined as overly noisy and inconsistent.

▼ What is preprocessing?

Preprocessing is a set of data mining techniques aimed at eliminating inconsistencies or inaccuracies in a dataset.