A Gentle Introduction to Model Selection for Machine Learning

by **Jason Brownlee** on December 2, 2019 in **Machine Learning Process**

Given easy-to-use machine learning libraries like scikit-learn and Keras, it is straightforward to fit many different machine learning models on a given predictive modeling dataset.

The challenge of applied machine learning, therefore, becomes how to choose among a range of different models that you can use for your problem.

Naively, you might believe that model performance is sufficient, but should you consider other concerns, such as how long the model takes to train or how easy it is to explain to project stakeholders. Their concerns become more pressing if a chosen model must be used operationally for months or years.
Also, what are you choosing exactly: just the algorithm used to fit the model or the entire data preparation and model fitting pipeline?

In this post, you will discover the challenge of model selection for machine learning.
After reading this post, you will know:

- Model selection is the process of choosing one among many candidate models for a predictive modeling problem.
- There may be many competing concerns when performing model selection beyond model performance, such as complexity, maintainability, and available resources.
- The two main classes of model selection techniques are probabilistic measures and resampling methods.

A Gentle Introduction to Model Selection for Machine Learning Photo by Bernard Spragg. NZ, some rights reserved.

Overview

This tutorial is divided into three parts; they are:

1. What Is Model Selection
2. Considerations for Model Selection
3. Model Selection Techniques

What Is Model Selection

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.

Model selection is a process that can be applied both across different types of models (e.g. logistic regression, SVM, KNN, etc.) and across models of the same type configured with different model hyperparameters (e.g. different kernels in an SVM).

When we have a variety of models of different complexity (e.g., linear or logistic regression models with different degree polynomials, or KNN classifiers with different values of K), how should we pick the right one?

— Page 22, Machine Learning: A Probabilistic Perspective, 2012.

For example, we may have a dataset for which we are interested in developing a classification or regression predictive model. We do not know beforehand as to which model will perform best on this problem, as it is unknowable. Therefore, we fit and evaluate a suite of different models on the problem.
Model selection is the process of choosing one of the models as the final model that addresses the problem.

Model selection is different from **model assessment**.
For example, we evaluate or assess candidate models in order to choose the best one, and this is model selection. Whereas once a model is chosen, it can be evaluated in order to communicate how well it is expected to perform in general; this is model assessment.

The process of evaluating a model's performance is known as model assessment, whereas the process of selecting the proper level of flexibility for a model is known as model selection.

— Page 175, An Introduction to Statistical Learning: with Applications in R, 2017.

Considerations for Model Selection

Fitting models is relatively straightforward, although selecting among them is the true challenge of applied machine learning.

Firstly, we need to get over the idea of a "*best*" model.

All models have some predictive error, given the statistical noise in the data, the incompleteness of the data sample, and the limitations of each different model type. Therefore, the notion of a perfect or best model is not useful. Instead, we must seek a model that is "*good enough*."

What do we care about when choosing a final model?

The project stakeholders may have specific requirements, such as maintainability and limited model complexity. As such, a model that has lower skill but is simpler and easier to understand may be preferred. Alternately, if model skill is prized above all other concerns, then the ability of the model to perform well on out-of-sample data will be preferred regardless of the computational complexity involved.

Therefore, a "*good enough*" model may refer to many things and is specific to your project, such as:

- A model that meets the requirements and constraints of project stakeholders.
- A model that is sufficiently skillful given the time and resources available.
- A model that is skillful as compared to naive models.
- A model that is skillful relative to other tested models.
- A model that is skillful relative to the state-of-the-art.

Next, we must consider what is being selected.

For example, we are not selecting a fit model, as all models will be discarded. This is because once we choose a model, we will fit a new final model on all available data and start using it to make predictions. Therefore, are we choosing among algorithms used to fit the models on the training dataset?

Some algorithms require specialized data preparation in order to best expose the structure of the problem to the learning algorithm. Therefore, we must go one step further and consider **model selection as the process of selecting among model development pipelines**.

Each pipeline may take in the same raw training dataset and outputs a model that can be evaluated in the same manner but may require different or overlapping computational steps, such as:

- Data filtering.
- Data transformation.
- Feature selection.
- Feature engineering.
- And more…

The closer you look at the challenge of model selection, the more nuance you will discover.

Now that we are familiar with some considerations involved in model selection, let's review some common methods for selecting a model.

Model Selection Techniques

The best approach to model selection requires "*sufficient*" data, which may be nearly infinite depending on the complexity of the problem.

In this ideal situation, we would split the data into training, validation, and test sets, then fit candidate models on the training set, evaluate and select them on the validation set, and report the performance of the final model on the test set.

If we are in a data-rich situation, the best approach […] is to randomly divide the dataset into three parts: a training set, a validation set, and a test set. The training set is used to fit the models; the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final chosen model.

— Page 222, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2017.
This is impractical on most predictive modeling problems given that we rarely have sufficient data, or are able to even judge what would be sufficient.

In many applications, however, the supply of data for training and testing will be limited, and in order to build good models, we wish to use as much of the available data as possible for training. However, if the validation set is small, it will give a relatively noisy estimate of predictive performance.

– Page 32, Pattern Recognition and Machine Learning, 2006.

Instead, there are two main classes of techniques to approximate the ideal case of model selection; they are:

- **Probabilistic Measures**: Choose a model via in-sample error and complexity.
- **Resampling Methods**: Choose a model via estimated out-of-sample error.

Let’s take a closer look at each in turn.

Probabilistic Measures

Probabilistic measures involve analytically scoring a candidate model using both its performance on the training dataset and the complexity of the model.

It is known that training error is optimistically biased, and therefore is not a good basis for choosing a model. The performance can be penalized based on how optimistic the training error is believed to be. This is typically achieved using algorithm-specific methods, often linear, that penalize the score based on the complexity of the model.

Historically various ‘information criteria’ have been proposed that attempt to correct for the bias of maximum likelihood by the addition of a penalty term to compensate for the over-fitting of more complex models.

– Page 33, Pattern Recognition and Machine Learning, 2006.

A model with fewer parameters is less complex, and because of this, is preferred because it is likely to generalize better on average.

Four commonly used probabilistic model selection measures include:

- Akaike Information Criterion (AIC).

- Bayesian Information Criterion (BIC).
- Minimum Description Length (MDL).
- Structural Risk Minimization (SRM).
-

Probabilistic measures are appropriate when using simpler linear models like linear regression or logistic regression where the calculating of model complexity penalty (e.g. in sample bias) is known and tractable.

Resampling Methods

Resampling methods seek to estimate the performance of a model (or more precisely, the model development process) on out-of-sample data.

This is achieved by splitting the training dataset into sub train and test sets, fitting a model on the sub train set, and evaluating it on the test set. This process may then be repeated multiple times and the mean performance across each trial is reported.

It is a type of Monte Carlo estimate of model performance on out-of-sample data, although each trial is not strictly independent as depending on the resampling method chosen, the same data may appear multiple times in different training datasets, or test datasets.

Three common resampling model selection methods include:

- Random train/test splits.
- Cross-Validation (k-fold, LOOCV, etc.).
- Bootstrap.

Most of the time probabilistic measures (described in the previous section) are not available, therefore resampling methods are used.

By far the most popular is the cross-validation family of methods that includes many subtypes.
Probably the simplest and most widely used method for estimating prediction error is cross-validation.

— Page 241, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2017.

An example is the widely used k-fold cross-validation that splits the training dataset into k folds where each example appears in a test set only once.

Another is the leave one out (LOOCV) where the test set is comprised of a single sample and each sample is given an opportunity to be the test set, requiring N (the number of samples in the training set) models to be constructed and evaluated.

Further Reading

This section provides more resources on the topic if you are looking to go deeper.

Tutorials

- Probabilistic Model Selection with AIC, BIC, and MDL

- A Gentle Introduction to Statistical Sampling and Resampling
- A Gentle Introduction to Monte Carlo Sampling for Probability
- A Gentle Introduction to k-fold Cross-Validation
- What is the Difference Between Test and Validation Datasets?

Books

- Applied Predictive Modeling, 2013.
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2017.
- An Introduction to Statistical Learning: with Applications in R, 2017.
- Pattern Recognition and Machine Learning, 2006.
- Machine Learning: A Probabilistic Perspective, 2012.

Articles

- Model selection, Wikipedia.

Summary

In this post, you discovered the challenge of model selection for machine learning.
Specifically, you learned:

- Model selection is the process of choosing one among many candidate models for a predictive modeling problem.
- There may be many competing concerns when performing model selection beyond model performance, such as complexity, maintainability, and available resources.
- The two main classes of model selection techniques are probabilistic measures and resampling methods.
- Do you have any questions?
- Ask your questions in the comments below and I will do my best to answer.