Data Science Toolkit

Matt Goodwin

October 23, 2018

Contents

1 Statistical Modeling

1.1 Overview and Theory

When discussing modeling it is important to keep in mind that "all models are wrong but some are useful" (attributed to George Box). The world is extremely complex and it can be impossible to create a model that perfectly approximates the underlying mechanisms that make our world turn.

There are different approaches to modeling depending on the discipline you come from, but personally I like the idea of the function approximation approach suggested by applied math and statistics. Taking this approach allows is to use probability theory combined with decision theory and to be able to visualize these concepts in a euclidean geometric space.

Bishop has a really nice overview of some of these concepts. The starting point I think for modeling starts with independent variable X (which could be a vector, see notation section) and dependent variable Y. We want to know:

1. The nature of the relationship between the variables (inference).
2. Given an independent variable, determine the dependent variable (prediction).

Using probability we can completely summarize the relationship and the uncertainty between the two variables with the joint distribution $P(X,Y)$. We use probability because in many problems we are interested in we generally cannot come to a completely deterministic relationship between the independent and dependent variables. This is partly because of measurement error, but also because the number of independent variables needed to perfectly determine the dependent variable is potentially infinite.

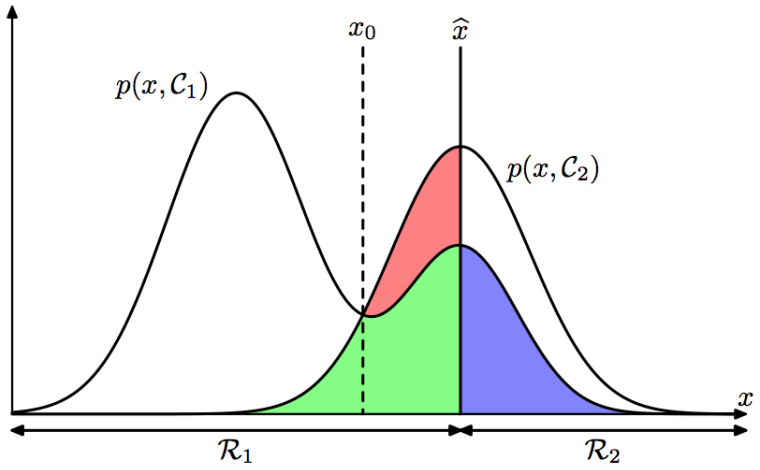
For example, imagine we wanted to predict the number of ice cream cones we sell on a particular day. Some variables such as the time of year or location of the ice cream store may provide us enough information to make a pretty good prediction or to understand the relationship between the independent and dependent variables fairly well. But to perfectly predict the number of ice cream cones we would need to know everything from the state of the road conditions, to whether or not a family from out-of-state decided to take a vacation. Since this is impossible we acknowledge variability and error in our estimates using probability.

I think the key to understanding this is to remember that the moment we use only a subset of all the possible features we would need for a perfectly deterministic relationship, then we must introduce uncertainty. We cannot say for certain that only knowing today is July 1 will lead to high ice-cream sales, but we can say the probability is higher than January 1st. When I have a training sample $(x_1,y_1),(x_2,y_2),...(x_n,y_n)$, I treat this as the truth (which it is) but I need to remember that these are draws coming from a distribution. I guess in that sense $P(X,Y)$ is a model itself, something we are forced to use because we don't know all the features needed for a deterministic relationship.

1.1.1 Decision Theory

As mentioned, we may want to perform inference, or in other words understand what $P(X,Y)$ looks like using information from a sample. This can give us an understanding of how the variables are related. In many practical applications however, we want to be able to predict Y given X. This is

Figure 1.1: Plot from Bishop showing visually the optimal decision boundary



where decision theory comes into play. Decision theory is designed to help us make the optimal decision given inputs. Bishop gives a nice overview that I try and summarize in my own words below.

Lets approach this by treating the dependent variable Y as a categorical variable taking on values 0 or 1. For simplicity assume X is a single continuous variable. We then have for $P(X,Y)$ a three-dimensional distribution where $P(Y|X)$ is a probability mass function. When making a decision called the *decision step* we formulate some rule that divides the input space into *decision regions*. If an instance falls into a certain decision region (based on X) it is predicted to be a 0 or 1. We want to minimize our mistakes as much as possible so we aren't assigning an instance to 0 when it should really be 1. The probability of a mistake can be written as:

$$P(mistake) = P(X \in R_1, 0) + P(X \in R_0, 1) \tag{1.1}$$

where R_1 is the region where an instance is assigned a 1 and R_0 is the region where an instance is assigned a 0.

Back to our example. Instead of ice cream sales, treat Y as a categorical variable where 1 is a good ice cream sales day and 0 is bad. If x_1 = "July 1st" is in R_1 we decide to assign it a 1, based on our decision rule. However, even though our model $P(X,Y)$ says that the probability of a high-selling day ($Y = 1$) is high in this region, there is a still a chance that it is a low-selling day because again, we are using a probability distribution for a model since we don't have all of the features we need for a deterministic model. The probability of it being a low-selling day for all X in R_1 is $P(X \in R_1, 0)$, which is a mistake.

We want to minimize our mistakes as much as possible so we choose regions where $P(X \in R_1, 0) + P(X \in R_0, 1)$ is as small as possible. To me it is easier to see this by thinking of the probability of being correct instead of the probability of being incorrect. This changes the problem from one of minimization to one of maximization. The optimal decision boundary therefore is the location that creates R_1 and R_0 such that $P(X \in R_1, 1) > P(X \in R_1, 0)$ everywhere in R_1 and $P(X \in R_0, 0) > P(X \in R_0, 1)$ everywhere in R_0. If the decision boundary were shifted either way then we would loose out on area under the distribution of being correct.

To visualize this better refer to figure 1.1.1 from Bishop. If our decision boundary were at x_0 then the

probability of being correct would be the two humped distribution completely colored in. This is the largest the probability of being correct can be. If we went with $\hat{x}$ however, then we loose out on the red region for being correct, which is suboptimal. (I like to think of a three dimensional distribution here whereas Bishop has an image with two different distributions which would need to be normalized appropriately but the concept is the same).

We can use the product rule to write:

$$\begin{aligned} P(X \in R_1, 1) > P(X \in R_1, 0) &= P(1|X \in R_1)P(X \in R_1) > P(0|X \in R_1)P(X \in R_1) \\ &= P(1|X \in R_1) > P(0|X \in R_1) \end{aligned} \tag{1.2}$$

So the maximization problem is equivalent to choosing the higher conditional probability for each region. This rule is known as the *Bayes classifier* and the error rate of the Bayes classifier is known as the *Bayes rate*. The Bayes classifier is used as a benchmark in classification as it is the optimal solution to classification if the probability distributions are known.

Maybe one of the most simplest models to start with is the additive error model. This is given simply as:

$$Y = f(X) + \epsilon \tag{1.3}$$

$$L(Y - \hat{f}(X)) = (Y - \hat{f}(X))^2 \mathbb{E}[L(Y - \hat{f}(X))] \tag{1.4}$$

1.2 Bias-Variance Tradeoff

The bias-variance tradeoff refers to two sources of error when evaluating models - the bias and the variance. There is also a third source of error which we call the "irreducible error".

As explained in this article, there is a slight confusion in data science between decomposing the error for an estimator, and decomposing the error for a model or a predictor. The decomposition is really about the same but there are some key insights to be aware of. The decomposition below is for a predictor. The decomposition for an estimator can be found in various books and other resources such as Casella/Berger.

First of all bias is defined as:

$$Bias() \tag{1.5}$$

1.3 Generalized Linear Models

There are three main components that make up the Generalized Linear Model (GLM):

1. Random component - assume the response variable comes from a probability distribution

$$Y_i \sim f(\mu_i) \tag{1.6}$$

where $\mu_i = E(Y_i)$ and f is a probability distribution.

2. Link component - connects the random component to the systematic component

$$g(\mu_i) = \eta_i \tag{1.7}$$

3. Systematic component - this is the linear part

$$\eta_i = x_i'\beta \tag{1.8}$$

1.3.1 Logistic Regression

Using the component concepts outlined above, for logistic regression we have:

1. Random component: $Y_i \sim f(\pi_i)$ where f is the Bernoulli distribution (since Y will be a binary variable when using logistic regression). $E(Y_i) = \pi_i$ which is the probability that Y_i is 1.

2. Link component: this is the logit function which is defined as:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \tag{1.9}$$

3. Systematic component: tying this all together we have:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i'\beta \tag{1.10}$$

Assumptions:

1. Linearity in log-odds

2. Independence of $Y_i|x_i$ and $Y_j|x_j$

3. Bernoulli response variable

For interpretation we can say that with a one unit increase in X_1 for example, then the log odds of a success goes up by β_1. We can also use the multiplicative odds where a one unit increase in X_1 leads to a e^{β_1} multiplicative change in odds on average.

The probability can be calculated by:

$$\begin{aligned}
\log\left(\frac{\pi_i}{1-\pi_1}\right) &= \beta_0 + x_i\beta_1 \\
\frac{\pi_i}{1-\pi_i} &= e^{\beta_0+x_i\beta_1} \\
\pi_i &= e^{\beta_0+x_i\beta_1} - e^{\beta_0+x_i\beta_1}\pi_i \\
\pi_i &= \frac{e^{\beta_0+x_i\beta_1}}{1+e^{\beta_0+x_i\beta_1}}
\end{aligned} \tag{1.11}$$

1.4 Boosting

The concept of boosting has lead to some of the most powerful algorithms in machine learning. Boosting falls under a general class of algorithms known as ensembles (bagging would be another

example of ensemble algorithms where we run separate models and then aggregate at the end by averaging for example). The general concept is that we run a weak learner on the original data, calculate the errors, run a new model on the errors, combine with the first weak learner, and repeat until some stopping criteria (that avoids overfitting).

1.4.1 ADABOOST

2 TERMS AND NOTATION

2.1 VARIABLE NOTIATION

Below explains notation used commonly when setting-up machine learning models. Note that all vectors are assumed to be column vectors. To help understand the notation I use the example of predicting the sales of ice cream cones.

- X - represents an input variable. Even though input variable implies a single variable this could also be a vector. If we wanted to access a single variable from the input vector then we use notation X_j. So for example X could include variables that describe the temperature (X_j), time or year (X_{j+1}), etc.
- Y - represents a *quantitative* output variable. This could be the sales of ice cream cones in dollars.
- G - represents a *qualitative* output variable. This could be if we sale over 50 ice cream cones for example (yes or no).
- x_i - represents an observed value of the variable X. Again this could be a vector. So to get the observed scalar value of the temperature for example we would write x_{ij}.
- $\boldsymbol{X}$ - matrix typically with dimensions Nxp.
- $\boldsymbol{x_j}$ - in general vectors are not bold unless the distinction is being made that this is the vector of all observation on X_j. So $\boldsymbol{x_j}$ is of length N and x_i is of length p.

3 BASIC STATISTICAL CONCEPTS

3.1 INFERENCE

Inference is referring to using data to figure out the underlying properties of a population (which in turn allows us to understand the relationship between variables). I've been thinking of inference as referring to the process to understand the relationship between variables in a linear regression, but I think this is too narrow of a view. For example, if we look at using a t-test to compare two samples what we are really doing is using the data to estimate what the two distributions are that the data comes from and then determine if that is reasonable or not. TODO: How do the various techniques in statistics fit in with this idea of finding the parameters of the underlying data?

GLOSSARY

dummy variable A vector where each element is either 0 or 1 and is used to represent a specific class. For example, if we have K classes then a dummy variable would be of length K and if we wanted to represent class 1, we would have a "1" in the first position in the vector and everywhere else would be 0.. 1

estimator A point estimator as defined by Cassella/Berger is any function $W(X_1, X_2, ..., X_n)$ of a sample. Any statistic is an estimator.. 1, 2

test A categorical variable that has ordering such as low, medium, and high, but no notion of a metric.. 1