MIT Big Data Challenge: Transportation in the City of Boston

Model of Prediction Challenge

Matt Edwards

My model used just the dropoff and pickup information. The event data was missing many important events, and I didn't want to use the MBTA data that was available for only half of the time period. I didn't have enough time to incorporate the weather or Twitter data. In general, my philosophy was that those extra data sources would be important for constructing and understanding a true causal model of taxi demand, but that observed taxi demand elsewhere would already incorporate those effects and be sufficient for predicting a small fraction of held-out points. That is, I could just leverage the correlation structure across time and space of observed taxi demand and not worry too much about exactly why things were happening (rain, a Celtics game, etc.). I think this strategy was decently successful, though incorporating other data sources (particularly Twitter density) might have reduced noise in my predictions.

My training examples were taxi pickups in the query time window for every day in the contest range. As candidate input features, I used pickups and dropoffs around the query zone, pickups and dropoffs before and after the query time, and pickups and dropoffs at key locations in the city. I used a variety of mostly linear models to train on these examples, including linear regression, the lasso, decision tree regression, and support vector regression. The regularization parameters for these models were all set by cross-validation. I used the initial test set answers to train a linear ensemble from all my separate model predictions, which I used to make the predictions for the final test set. The models I used all minimized squared error, which was the scoring metric the contest used. This was more of a lucky coincidence than great planning, since squared error has nice computational properties. I tried other techniques that are more technically appropriate for count data, but they didn't improve the results much.

I don't think my model would be very useful to city planners, due to the structure of the prediction task. We are predicting observed or achieved taxi demand, not desired or optimal taxi demand. Our training set and models only include trips where a taxi was available and picked up an individual. There is no notion of wait time for a taxi or unfulfilled taxi demand in the dataset that would necessitate adding another cab stand. If an individual gives up on waiting for a taxi or walks several blocks for a better hailing location, we are blind to this fact. It may be inferred indirectly if there are many pickups from a location that is not a cab stand, but a better option would be compiling a dataset that includes wait times before getting into a cab. This might involve dispatcher data and the times (and distances) between calls and pickups. From a technical perspective, my modeling approach would work and make predictions without incorporating present or future information, but the results would certainly have been less accurate.