Sean Harnett

3-26-12

APAM E4990 HW #2

1 Cross-validation for polynomial regression

The optimal degree polynomial varied depending on the particular 50/50 split used. Degree four was best most often in 100 trials, see figure 1. For one split that resulted in degree four being optimal, the polynomial coefficients were:

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} 0.9872 \\ 0.2221 \\ -1.7407 \\ 0.3502 \\ 0.8884 \end{bmatrix}$$

See figure 2 for a plot of error as a function of polynomial degree for this split, as well as a scatter plot of the data with the fourth degree polynomial overlayed.

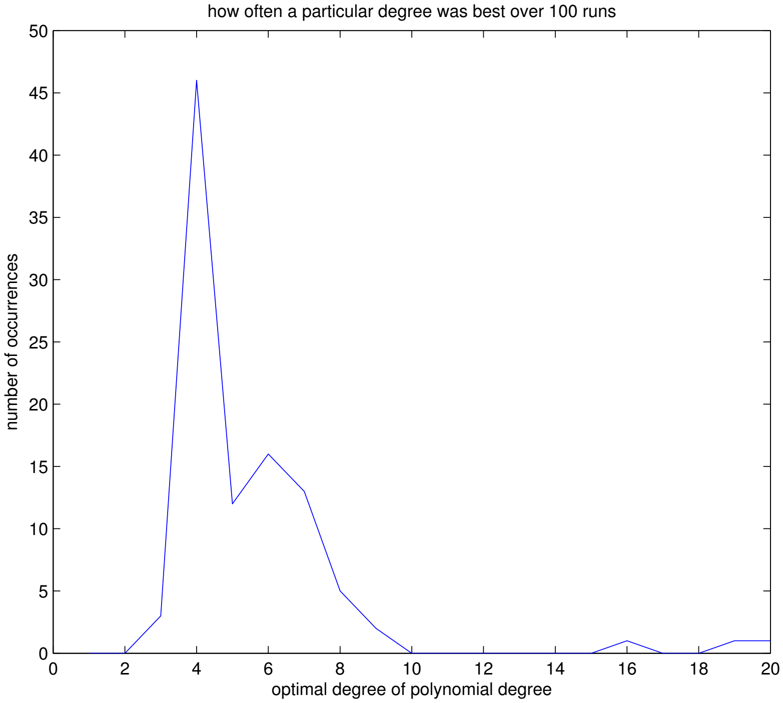

Figure 1: 100 random 50/50 splits of training and test data, and the number of times each degree polynomial was optimal.

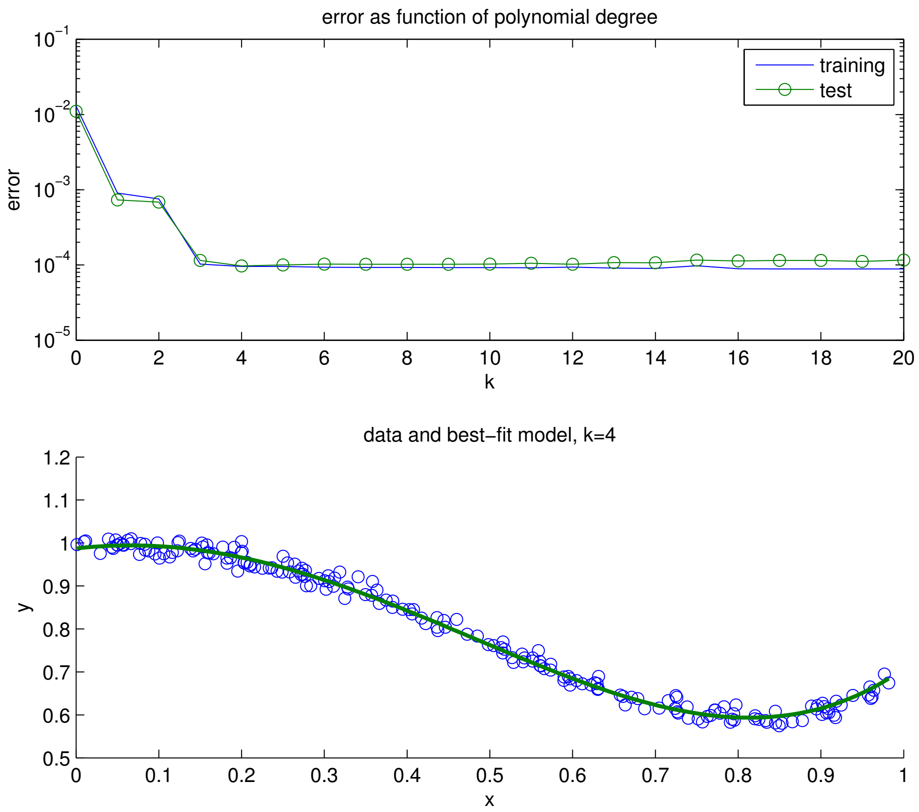

Figure 2: Error as a function of polynomial degree, and scatter plot of data with best polynomial.

2 Naive Bayes for article categorization

Performance

download_articles.py

5 minutes 38 seconds, two errors (no body in article). This uses the nytimes API to get the latest 2000 articles from each of the five categories. It creates a tab-delimited file for each category: each line is an article, each article has three fields: url, title, body.

parse_features.py

1.6 seconds. This creates the list of words, and the data structure used by the naive bayes algorithm. It opens up each tab-delimited file, and looks at the set of words in the title and body. It removes punctuation and ignores stop words. It pickles two data structures: a list of every word in all the articles, and a list of articles, each of which is a list of indices corresponding to words in the overall list.

train.py

$\approx$ 2.5 to 4 seconds per run, depending on how much stuff you want. This actually does Naive Bayes. Producing just the confusion table and test error for a single α, β pair takes about 2.5 seconds. For the 50/50 split using seed(0), and α, $\beta = 1.5$, it achieves the following accuracies:

training accuracy: .9908

test accuracy: .9203

Effects of changing α and β

Running "python train.py optimize" will try out a grid (log-scale) of α's and β's. The accuracy was much more sensitive to changes in α; increasing beyond $\alpha > 3$ or so resulted in much poorer performance, with a sweet spot somewhere near 1.5. Changes in β had a more subtle effect, and values as large as 100 would give good results for certain α's. Using classification error (as opposed to something smooth) made optimizing these a bit inexact; most choices in the neighborhood of α, $\beta = 1.5$ gave about the same results.

Confusion table

This is for α, $\beta = 1.5$ on the test set, where the accuracy was .9203.

		Predicted				
		Arts	Business	Obituaries	Sports	World
Actual	Arts	913	30	28	18	26
	Business	33	917	8	5	30
	Obituaries	14	3	889	20	28
	Sports	13	15	8	978	10
	World	19	78	10	1	885

Top 10 most informative words for each section

Nothing too surprising here. "Arts" is the base category, so for the four other categories, these are simply the words with the most weight, i.e. $\text{argmax}_j w_{jc}$. For Arts, these are the words with the least weight (big negative number) over the four other categories. Not surprisingly, some of these words are the same for different categories, so I'm showing the top ten unique ones.

Arts	Business	Obituaries	Sports	World
music	dealbook	died	briefing	briefing
books	profit	dies	roundup	officials
artsbeat	bonds	complications	ncaa	killed
theater	stocks	confirmed	mens	iran
2012	tax	88	sports	syrian
movie	quarter	86	bowl	attack
review	euro	87	nfl	kabul
sales	rate	90	round	opposition
lists	consumer	89	visiting	beijing
print	bits	82	nhl	election

Top 10 “most difficult to classify" articles

I played with a couple definitions of “difficult to classify”, but they all came up with the same sorts of articles: ones that should have multiple categories (which I tried to remove during preprocessing), or that are simply miscategorized by the API. I settled on the definition “misclassified with large difference between top two most likely categories”. This definition picks out articles where the classifier is very confident of its prediction, yet is wrong. The top mistake with this definition was the following article:

> Wall Street Ahead Strongly After Retail Data - Stocks on Wall Street traded strongly higher Tuesday after the Commerce Department reported better-than-expected retail sales in the United States last month despite rising gas prices. The Standard & Poor’s 500-stock index gained 1.4 percent in late trading. The Dow Jones industrial average added 1.3 percent and the Nasdaq composite index rose 1.5

The classifier predicted *Business* for this article, which not only looks accurate, but also matches the URL:

> http://www.nytimes.com/2012/03/14/business/daily-stock-market-activity.html

This article came up under *World* from the API. Another common error was an article that should probably be in two categories, like the following article, reporting the death of someone in the arts:

> One of the highly regarded veterans of tournament bridge, Russ Arnold of Miami, died on Jan. 27 at 90 after a prolonged battle with prostate cancer and kidney failure. The death was announced in an e-mail by Julie and Walter Murphy of Hendersonville, N.C., who were very close friends of Arnold’s. He won the 1981 Bermuda Bowl world championship and

A few articles came up in multiple categories; the code removes these, so that isn’t the problem here. Finally, a number of articles were just tricky, and could probably fool a human who wasn’t paying careful attention. For instance, this article on a bridge tournament (Arts) sounds awfully like a sports article:

> On Jan. 16, the last day of the District 3 Winter Regional in Rye Brook, N.Y., a nine-table youth pairs event included 20 players who arrived on a bus from New Jersey. The winners were Amber Yu Lin, 14, and Jennifer Ling, 15, of Edison, N.J. They had a 78.78 percent game to finish one board ahead of Brandon Lin (Amber’s 12-year-old brother) and

The full top ten is at the end.

How classifier might generalize to other contexts

e.g. articles from other sources or time periods. My guess is it would generalize poorly. Other sources (from other countries, say) and other time periods likely have very different concerns. For example, they might be interested in entirely different sports and art forms. Life expectancies could be different, so famous people might die younger. “World” would certainly have a different definition than nytimes has.

Of course the closer the context to that of nytimes, the better this classifier would perform.

Full list of tricky articles

```
actual: World predicted: Business
http://www.nytimes.com/2012/03/14/business/daily-stock-market-activity.html
Wall Street Ahead Strongly After Retail Data
Stocks on Wall Street traded strongly higher Tuesday after the Commerce Department
reported better-than-expected retail sales in the United States last month despite
rising gas prices. The Standard & Poor’s 500-stock index gained 1.4 percent in late
trading. The Dow Jones industrial average added 1.3 percent and the Nasdaq
composite index rose 1.5
```

```
actual: World predicted: Business
http://www.nytimes.com/2012/02/19/technology/foxconn-to-raise-salaries-for-workers-b
y-up-to-25.html
Foxconn Plans To Sharply Lift Workers’ Pay
BEIJING -- Foxconn Technology, one of the biggest manufacturers of products for
Apple, Dell, Hewlett-Packard and other electronics companies, said Saturday that it
would sharply raise worker salaries at its Chinese factories. Foxconn said that
salaries for many workers would immediately jump by 16 to 25 percent, to about $400
a month, before
```

```
actual: Arts predicted: World
http://www.nytimes.com/2012/03/18/books/review/david-c-ungers-the-emergency-state.html
Fear Factor
THE EMERGENCY STATE America’s Pursuit of Absolute Security at All Costs By David C.
Unger 359 pp. The Penguin Press. $27.95. With Osama bin Laden dead, American troops
leaving Iraq, the economy still sputtering and Congress locked in yet another
budget showdown, one thing that seems clear is that Washington will very likely cut
military spending
```

```
actual: Arts predicted: Sports
http://www.nytimes.com/2012/01/28/crosswords/bridge/district-3-winter-regional-hosts
-young-bridge-players.html
BRIDGE; In Youth Pairs Event, 2 Girls Win a Big Game
On Jan. 16, the last day of the District 3 Winter Regional in Rye Brook, N.Y., a
nine-table youth pairs event included 20 players who arrived on a bus from New
Jersey. The winners were Amber Yu Lin, 14, and Jennifer Ling, 15, of Edison, N.J.
They had a 78.78 percent game to finish one board ahead of Brandon Lin (Amber’s
12-year-old brother) and
```

```
actual: Arts predicted: Obituaries
http://www.nytimes.com/2012/02/02/crosswords/bridge/russ-arnolds-defense-at-bermuda-
bowl-semifinal-bridge.html
BRIDGE; Remembering Deft Defense Of a Hall of Fame Champion
One of the highly regarded veterans of tournament bridge, Russ Arnold of Miami,
died on Jan. 27 at 90 after a prolonged battle with prostate cancer and kidney
failure. The death was announced in an e-mail by Julie and Walter Murphy of
Hendersonville, N.C., who were very close friends of Arnold’s. He won the 1981
Bermuda Bowl world championship and
```

```
actual: Arts predicted: Sports
http://query.nytimes.com/gst/fullpage.html?res=9503E4DE103DF933A25751C0A9649D8B63
What’s On Today 7 P.M. (USA) NFL CHARACTERS UNITE
Tony Gonzalez of the Atlanta Falcons, Jimmy Graham of the New Orleans Saints, Hines
Ward of the Pittsburgh Steelers (above right, with Carlton Dennis, a student) and
```

```
Tony Dungy, a former head coach for the Indianapolis Colts and the Tampa Bay
Buccaneers, discuss the hate and bigotry they overcame as children to

actual: World predicted: Business
http://www.nytimes.com/2012/02/21/world/europe/agreement-close-on-a-bailout-for-gree
ce-european-finance-ministers-say.html
Europe Agrees on New Bailout To Help Greece Avoid Default
BRUSSELS -- Greece finally secured its second giant bailout early Tuesday after
euro zone finance ministers agreed to save it from bankruptcy in exchange for
severe austerity measures and strict conditions. After more than 13 hours of talks,
the ministers approved a new bailout of 130 billion euros, or $172 billion, under
which private investors in

actual: Business predicted: Sports
http://www.nytimes.com/2012/03/15/business/media/luring-stanley-cup-viewers-with-a-t
orrent-of-games.html
ADVERTISING; Luring Stanley Cup Viewers With a Torrent of Games
TURNING the Stanley Cup playoffs into a must-watch television event lasting for
weeks in the spring is no easy feat. Many Americans did not grow up with the game,
few know who the players are and watching 10 weeks of playoffs is a big time
commitment. But the National Hockey League’s plan to get more viewers is, in fact,
to show more games. This

actual: Obituaries predicted: Sports
http://www.nytimes.com/2011/06/29/sports/ncaabasketball/lorenzo-charles-47-made-winn
ing-dunk-in-1983-ncaa-title-game.html
Lorenzo Charles, 47; Dunk Won 1983 Title
Lorenzo Charles, whose dunk in the final seconds of the 1983 National Collegiate
Athletic Association national championship game propelled North Carolina State
University to victory over Houston and himself to the realm of basketball legend,
died Monday when the charter bus he was driving crashed in Raleigh, N.C. He was 47.
North Carolina State

actual: World predicted: Business
http://www.nytimes.com/2012/03/16/world/asia/bain-capital-tied-to-surveillance-push-
in-china.html
A U.S. tie to Push On Surveillance In Chinese Cities
BEIJING -- As the Chinese government forges ahead on a multibillion-dollar effort
to blanket the country with surveillance cameras, one American company stands to
profit: Bain Capital, the private equity firm founded by Mitt Romney. In December,
a Bain-run fund in which a Romney family blind trust has holdings purchased the
video surveillance
```