Digging to Data: Homework 3
Due: 23:59, --

The goal of this homework is to predict whether the answer to a trivia question is correct or not. Thus, it is a binary classification problem. You will have to train effective classification models, but the more important (and effective) route for success is to engineer features and gather additional data to help your predictions.

You will need to provide four deliverables for this assignment: your code for completing this assignment, a writeup detailing your steps (answering the steps highlighted in bold), a single plot that shows a new feature you added, and finally predictions on a set of test data. You'll also (as last time) be competing with your classmates on Kaggle to make good predictions.

About the Data

Quiz bowl is an academic competition between schools in English-speaking countries; hundreds of teams compete in dozens of tournaments each year. Quiz bowl is different from Jeopardy, a recent application area. While Jeopardy also uses signaling devices, these are only usable *after a question is completed* (interrupting Jeopardy's questions would make for bad television). Thus, Jeopardy is rapacious classification followed by a race---among those who know the answer---to punch a button first.

Here's an example of a quiz bowl question:

> Expanding on a 1908 paper by Smoluchowski, he derived a formula for the intensity of scattered light in media fluctuating densities that reduces to Rayleigh's law for ideal gases in *The Theory of the Opalescence of Homogenous Fluids and Liquid Mixtures near the Critical State*. That research supported his theories of matter first developed when he calculated the diffusion constant in terms of fundamental parameters of the particles of a gas undergoing Brownian Motion. In that same year, 1905, he also published *On a Heuristic Point of View Concerning the Production and Transformation of Light*. That explication of the photoelectric effect won him 1921 Nobel in Physics. For ten points, name this German physicist best known for his theory of Relativity.
>
> ANSWER: Albert **Einstein**

Two teams listen to the same question. Teams interrupt the question at any point by "buzzing in"; if the answer is correct, the team gets points and the next question is read. Otherwise, the team loses points and the other team can answer.

Why we want to use Quiz Bowl Data for Classification

It's very easy to generate guesses (in fact, we could generate every possible guess). The challenge is knowing whether any given guess is good or not. We can treat this as a classification problem. Every guess can be described by features that measure how well it matches the question. The classifier tells us whether we got the question wrong or right.

Data Format

Each line has a guess (page) and a correct answer (answer) given some fraction of the question revealed so far (text). Your goal is to predict whether they match. Each guess is the title of a Wikipedia page. To get you started, you have the following columns

- **row** A unique ID for every guess
- **body_score** A measure of the confidence for how well the text of the Wikipedia page matches the question text
- **page** The guess produced
- **answer** The correct answer
- **text** The text so far revealed (randomly chosen)
- **category** The category (could be noisy) of the question
- **tournaments** The tournaments the question was used in
- **answer_type** The kind of answer of the question (e.g. person, place, work)
- **corr** Whether the guess (page) was correct or not (i.e., whether it matched "answer")
- **inlinks** The number of inlinks on the guessed Wikipedia page

---

1. Download from the Kaggle site (link to be provided.
    - *qb.train.csv* - the training set
    - *qb.test.csv* - the test set (missing the column "corr")
    - *qb.guess.csv* - a sample submission file in the correct format
    - *feature_expansion.R* - A script to run svm classification and extract additional features
2. (30 points) Build the best classifier you can with the given data, documenting the choices that you make.
    a. **Try using logistic regression, SVM (multiple kernels), and decision trees. Create a table with your accuracy with each of these methods.**
    b. **Look at where you're making mistakes. Can you see any patterns?**
3. (40 points) Find additional information you can use to improve predictions. Be creative. Look for features you can extract from the data that you have. NOTE: To get credit for this, you need to have an idea and evaluate it. You will get full credit for a well-thought out feature that doesn't improve performance.
    a. **Create a plot that explains why this information is useful for making predictions (e.g. create a facet graph showing the distribution for correct**

and incorrect). Turn this in as "plot.pdf".

b. **How much does this feature improve your classification?**

4. (30 points) Challenge: Build a classifier that best predicts correct answers in this dataset. Upload your predictions to Kaggle. (note that you must use your UMD e-mail to get access to this competition). If your UMD username is JSMITH, use JSMITH_DID as your username on Kaggle. You are welcome to use any additional data you care to use so long as they are not from quiz bowl questions.
   a. **Provide your final score and username**
   b. **Create an error analysis of your final classifier. Turn this in as "error.pdf". An error analysis must contain real examples of your data, not just an error matrix.**

      A good error analysis must contain examples from the development set that you get wrong. You should show those sentences and explain why (in terms of features or the model) they have the wrong answer.

      You should have been doing this all along as your derive new features (e.g., 2b), but this is your final inspection of your errors. The feature or model problems you discover should not be trivial features you could add easily. Instead, these should be features or models that are difficult to correct.

      An error analysis is not the same thing as simply presenting the error matrix, as it does not inspect any individual examples.