STAT 745 – Fall 2014
Assignment 1

Group 1: Francene Cicia, Lonie Moore, Doug Raffle, Melissa Smith

August 25, 2014

1 Artificial Intelligence vs. Machine Learning

1.a Identify and describe $\mathcal{G}, y, x, A$

- $\mathcal{G}$ is the set of possible response classes. In this case, there are $k = |\mathcal{G}| = 2$ levels: Artificial Intelligence (AI) and Machine Learning (ML) (i.e., $\mathcal{G} = \{\text{AI, ML}\}$).
- y is the vector of responses for the $n = 879$ observed papers ($y_i \in \mathcal{G}$).
- x is a 879×141 matrix showing the count of the tokens in the title of every paper (where $p = 141$ is the number of unique tokens in the corpus). For every row x_i, x_{ij} is the frequency of token j in paper i's title.
- A is a 879×879 symmetrical matrix which represents a graph of the papers' shared citations. Each A_{ij} is the number of citations that papers i and j share. The diagonals, A_{ii}, are the number of citations for paper i.

1.b Strategies

If we were to view this as a supervised problem, we would treat the topic (y) as a response and train a model to classify papers based on the words in the title and/or the citation network, using a technique like K-Nearest Neighbors.

In an unsupervised context, we would ignore the class (or treat it as another feature) and use some form of Cluster Analysis to see how similar the observations are to each other, though there is no guarantee that the papers would cluster by topic.

1.c The First Paper

The tokens in the title are:

```
> inds <- which(tokens[1,] > 0)
> cat(names(tokens[1,inds]), "\n", sep="\t")

experiment      induction       plan        with
```

Since the titles were tokenized as lexemes (i.e., word roots), not whole words, and most function words (e.g., prepositions, articles, etc.) have been discarded, there is obviously some guesswork. The title may be something like "Planning Experiments with Induction," "Experimental Planning with Induction," or "Induction with Experimental Planning." It's impossible to be at all certain without inflectional prefixes or suffixes, function words, or more context.

A simple method for finding the papers most similar to the first would be to treat the papers as points in p-space, and selecting the papers that are the closest to the first using Euclidean Distances.

```
> library(parallel); cl <- makeCluster(detectCores()); clusterExport(cl, "tokens")
> dis <- parRapply(cl, tokens[-1,], function(r) sqrt(sum((tokens[1,]-r)^2)))
> stopCluster(cl)
> (nearest.tok <- as.integer(names(head((sort(dis))))))

[1]  37 116 120 217 368 538
```

We would hope, then, that these papers share the same topic, at least on average, as the first (y_1 =AI):

```
> as.character(class[nearest.tok])

[1] "ML" "AI" "AI" "AI" "AI" "ML"
```

This method, however, only considers the token features. To find similar papers in terms of the citation graph, we need only find the papers (nodes) that have highest number of shared citations (edge values).

```
> names(cite) <- gsub("node_", "", names(cite))
> (nearest.g <- as.integer(names(head(sort(unlist(cite[1, -1]), decreasing=TRUE)))))

[1]   2   3 574 360 576 626

> as.character(class[nearest.g])

[1] "AI" "AI" "AI" "AI" "AI" "AI"
```

The graph based approach seems to outperform the tokenized titles in selecting papers with the same topic.

2 Absolute Error Loss

$$\hat{f}(x) = \arg\min_c \ \mathrm{E}_{Y|X}\left[|Y - c| \mid X = x\right]$$

We begin by re-writing the expectation:

$$\mathrm{E}_{Y|X}\left[|Y - c| \mid X = x\right] = \int_{-\infty}^{\infty} |y - c| \ f_x(y)\ dy = \int_{-\infty}^{c} (c - y)\ \ f_x(y)\ dy + \int_{c}^{\infty} (y - c)\ \ f_x(y)\ dy$$

Then we take the derivative using Leibniz's Rule:

$$\begin{aligned}\frac{\partial}{\partial c}\int_{-\infty}^{c} (c - y)\ \ f_x(y)\ dy + \frac{\partial}{\partial c}\int_{c}^{\infty} (y - c)\ \ f_x(y)\ dy &= \int_{-\infty}^{c} \frac{\partial}{\partial c}(c - y)\ \ f_x(y)\ dy + \int_{c}^{\infty} \frac{\partial}{\partial c}(y - c)\ \ f_x(y)\ dy \\ &= \int_{-\infty}^{c} f_x(y)\ dy - \int_{c}^{\infty} f_x(y) dy\end{aligned}$$

Which we set to zero and solve for c to find $\arg\min_c$:

$$\begin{aligned}\int_{-\infty}^{c} f_x(y)\ dy - \int_{c}^{\infty} f_x(y) dy &= 0 \\ \int_{-\infty}^{c} f_x(y)\ dy &= \int_{c}^{\infty} f_x(y) dy \\ Pr\left(Y \leq c \mid X = x\right) &= Pr\left(Y > c \mid X = x\right) \\ Pr\left(Y \leq c \mid X = x\right) &= 1 - Pr\left(Y \leq c \mid X = x\right) \\ 2Pr\left(Y \leq c \mid X = x\right) &= 1 \\ Pr\left(Y \leq c \mid X = x\right) &= \frac{1}{2}\end{aligned}$$

So $\arg\min_c$ is, by definition, the conditional median of Y given $X = x$, leaving us with:

$$\hat{f}(x) = \arg\min_c \ \mathrm{E}_{Y|X}\left[|Y - c| \mid X = x\right] = \mathrm{median}\left(Y \mid X = x\right)$$

Which tells us that using L_1 loss for K-NN is to quantile regression what using L_2 loss is to least squares regression.