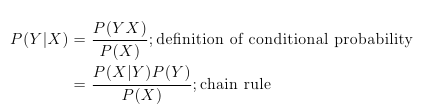
(Faiyam Rahman, Alan Hau

Modern Analytics

HW 3

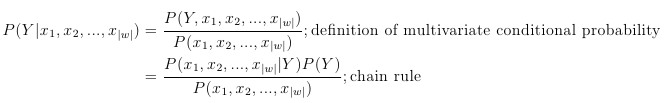
12/1/14

**Question 1**

**1A) **

**1B)** The equation in 1a is always true, as the only two propositions invoked are the definition of conditional probability, which is always true by definition, and the chain rule, which is always true for arbitrary random variables X, Y.

**1C)**

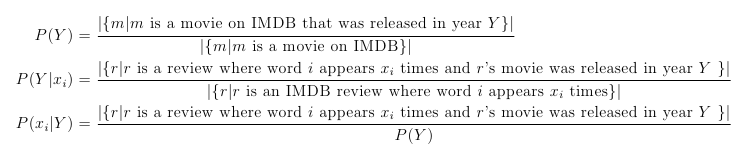


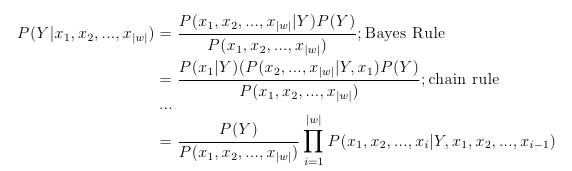
**1D)** P(Y) is the probability that a given movie listed on IMDB belongs to year Y.

P(Y|Xi) is the probability that a given movie listed in IMBD belongs to year Y given that word Wi appears Xi  times in the movie’s IMDB review.

P(Xi|Y) is the probability that word Wi appears Xi times in a given movie’s IMDB review, given that the movie was released in year Y.

Equations to calculate each of the above quantities are as follows:



**1E) **

**1F)** The equation in 1a is always true, as the only two propositions invoked are Bayes rule and the chain rule, both of which are always true for arbitrary multivariate distributions.

**1G)** Here’s a short bit of pseudo-code that would do the job. I understand why it’d be hard to actually do.

*R = set of all reviews*

*function calcProbability(Y, x1, x2, … , xn):*

*let count\_num = 0*

*for review in R:*

*if (word 1 appears in x1 times) and*

*(word 2 appears x2 times) and*

*…*

*(word n appears xn times) and*

*(the review is for a movie in year Y):*

*count\_num += 1*

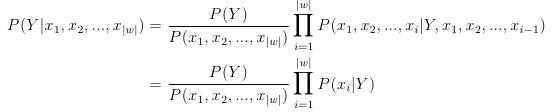
*return count\_num/size(R)*

*print calcProbability(Y, x1, x2, …, xn)/calcProbability(Y, x1, x2, x3, …, x\_(n-1))*

**1H)** The “Naïve Bayes” assumption, in a general sense, says that for a feature vector **X** = (X1,X2, … , Xn), we assume that Xi is independent of Xj for all i, j such that i != j. In our specific case, it means that two words Wi and Wj, the number of times word Wi appears in a given IMDB review r is independent of the number of times word Wj appears in review r.

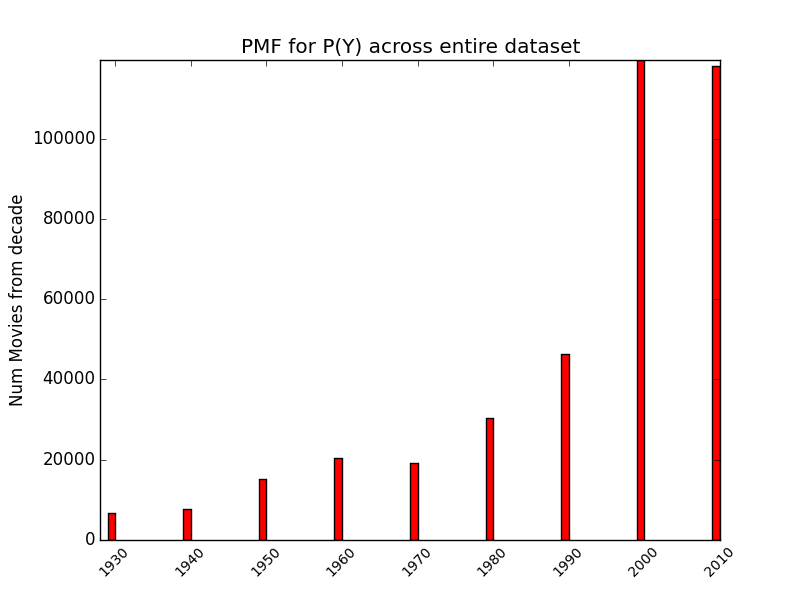
In our specific case, it is violated in many situations. For example, the word “the” is more likely to appear in a given review r if the word “to” is in review r, due to the fact that people write using common English grammar (e.g: I went to the movie).

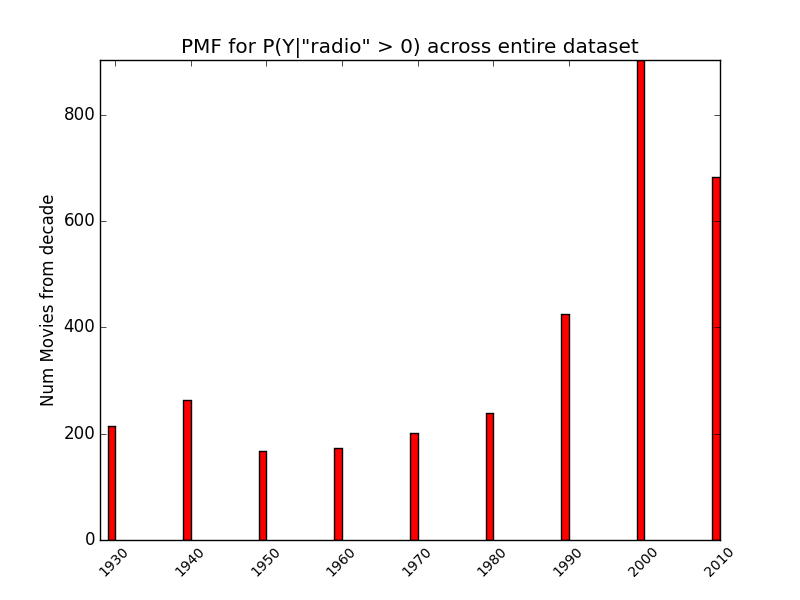
**1I)** This vastly simplifies our equation for P(Y|X1, X2, … , Xn).

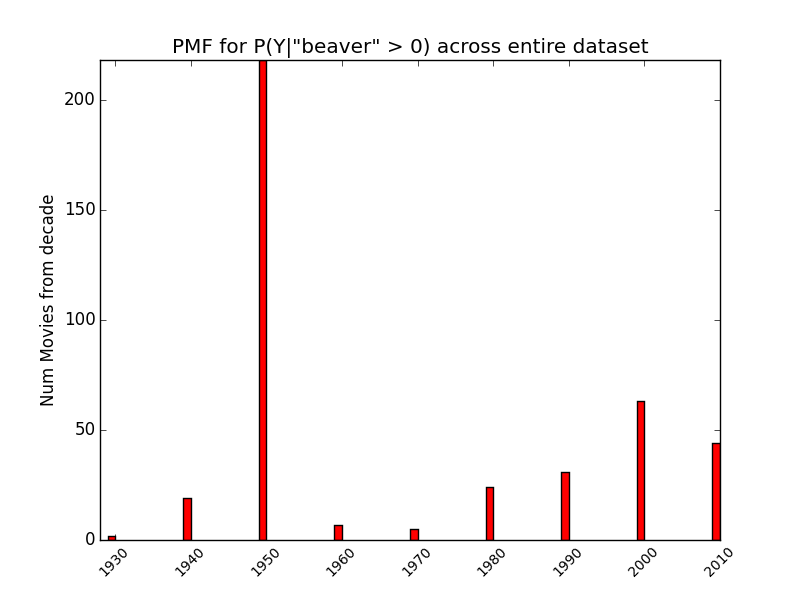


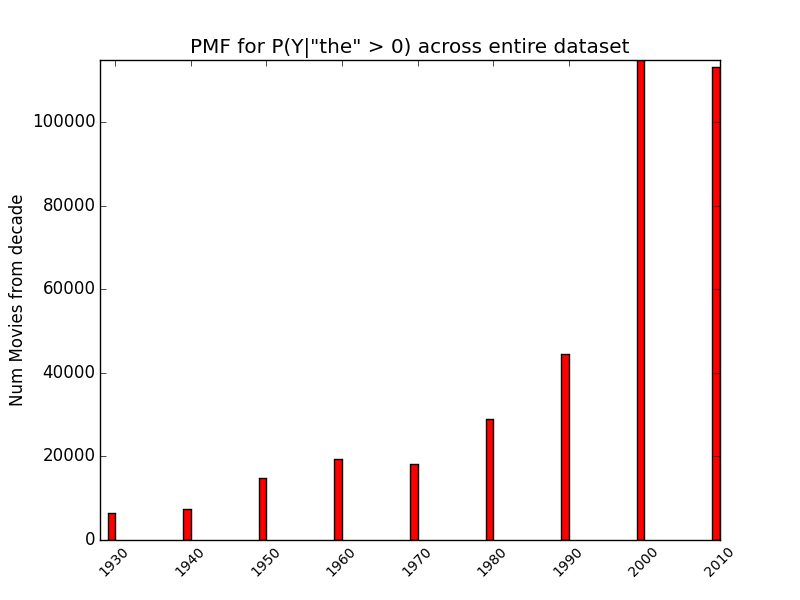
**Question 2**

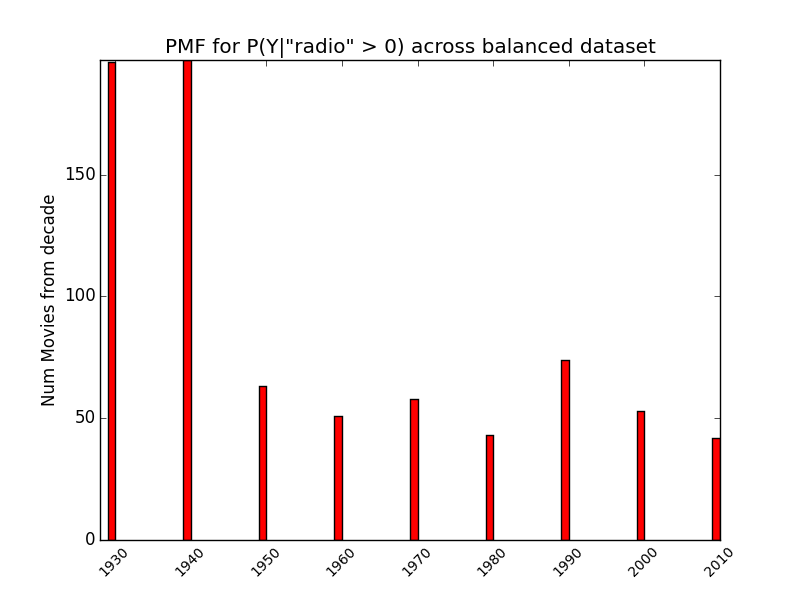
**2a-g)** Please see below for the plots for 2a through 2g

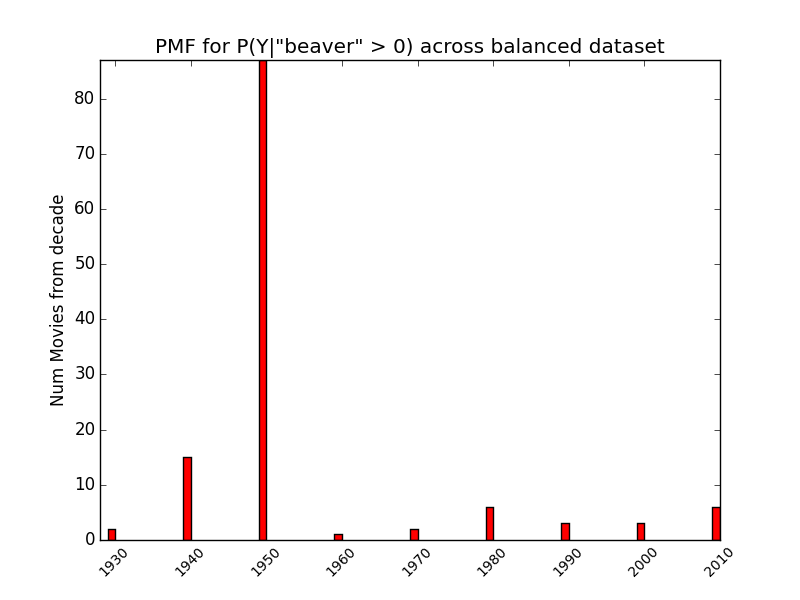


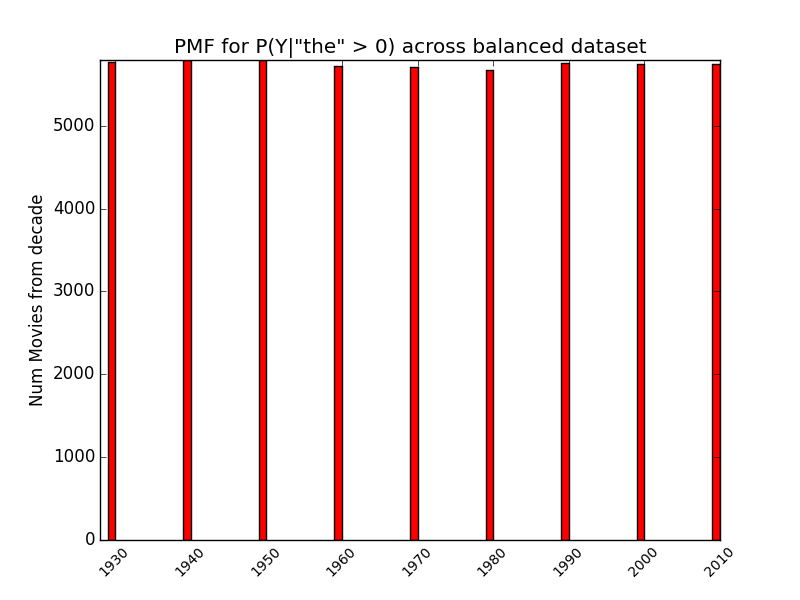






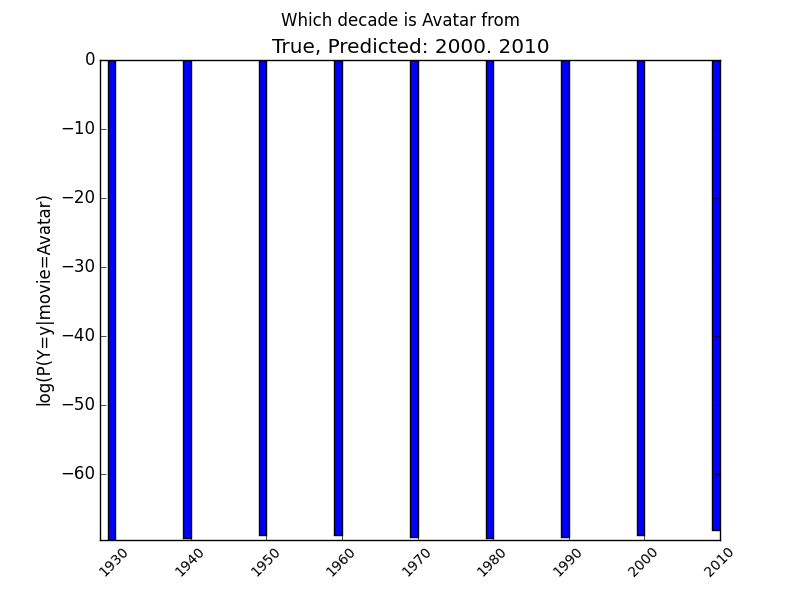


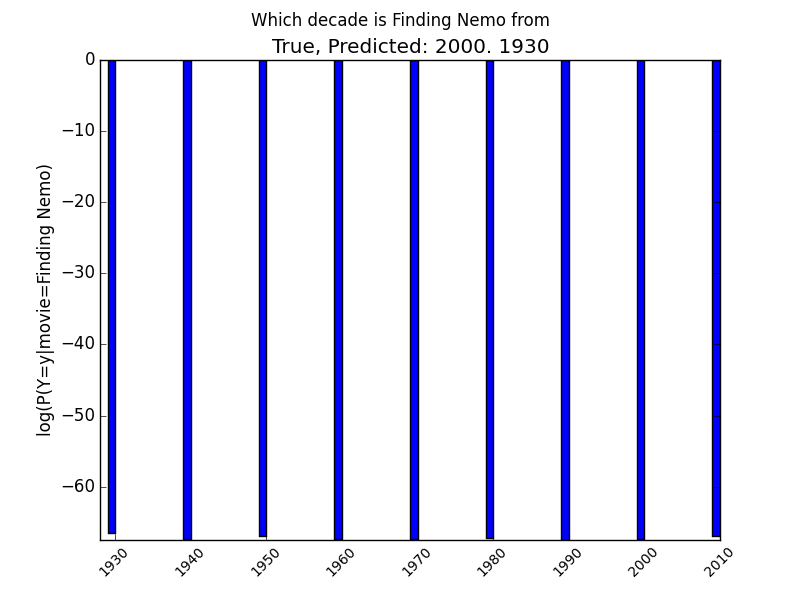


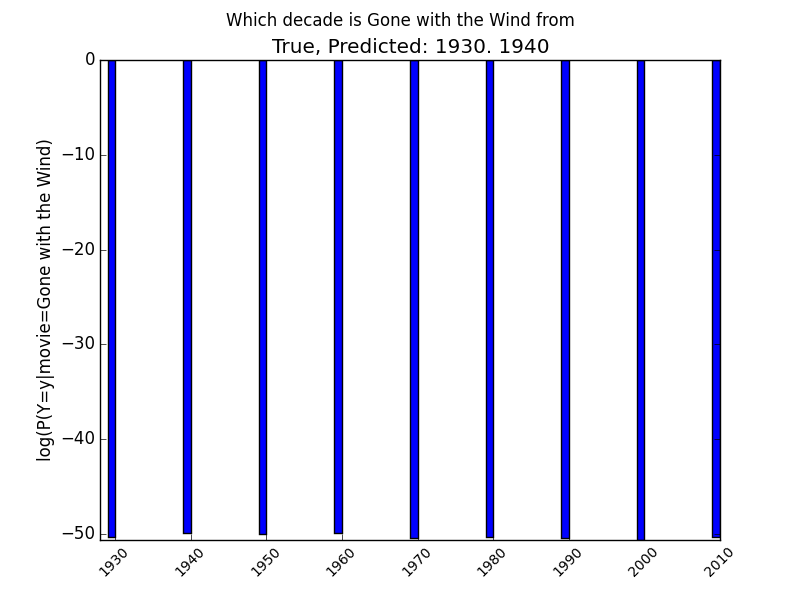


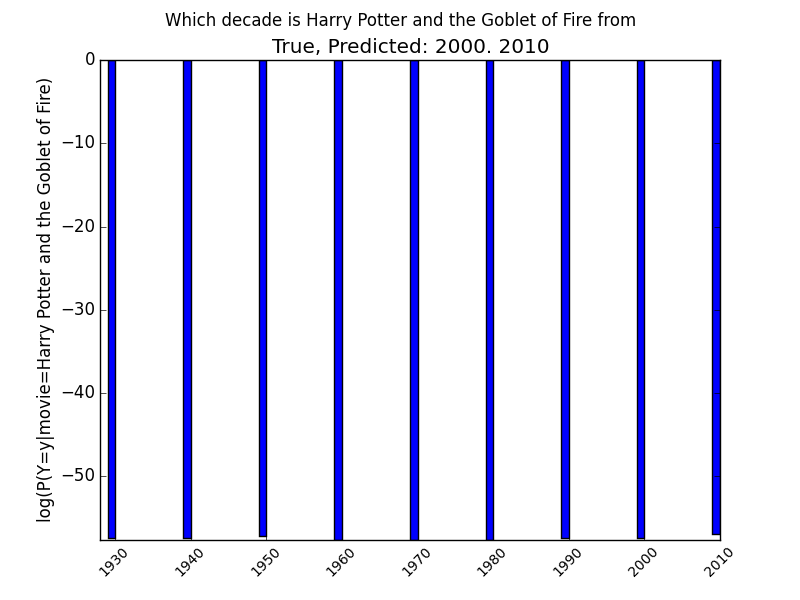
**2H)** It’s much easier to compare probabilities across years in the PMFS over the balanced dataset. Balancing the dataset made the PMFs reflect probabilities, not the prior distribution of decades in the dataset.

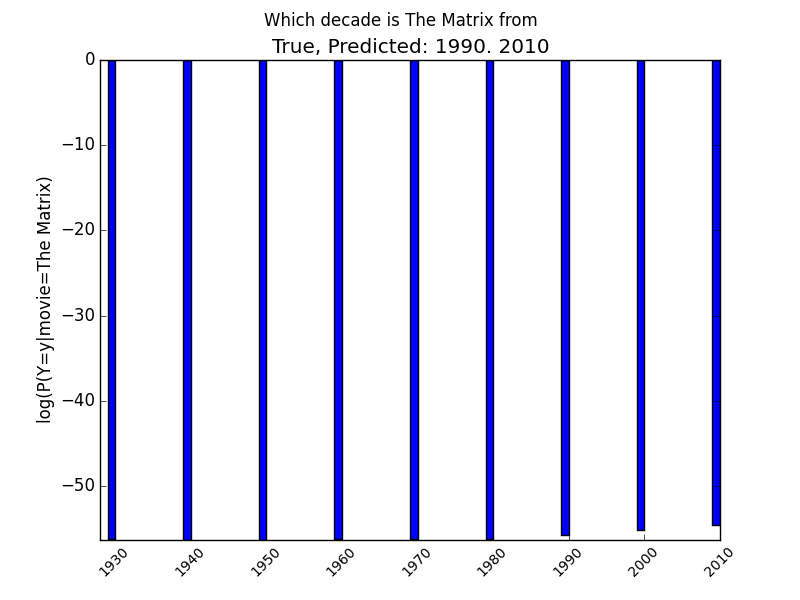
**2J)** Please see below for the plots**.** Our classifier failed to correctly classify any of the 5 movies, but was within +/- 1 decade of 4 of the movies.







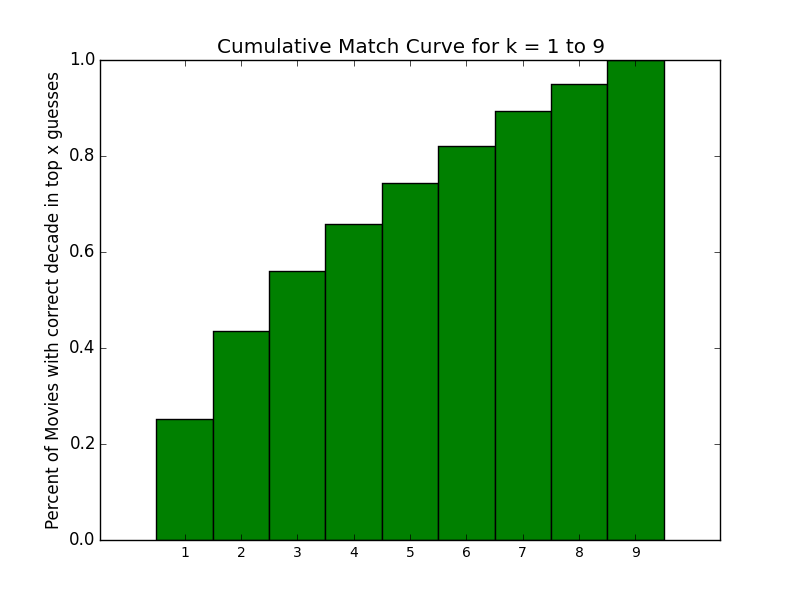




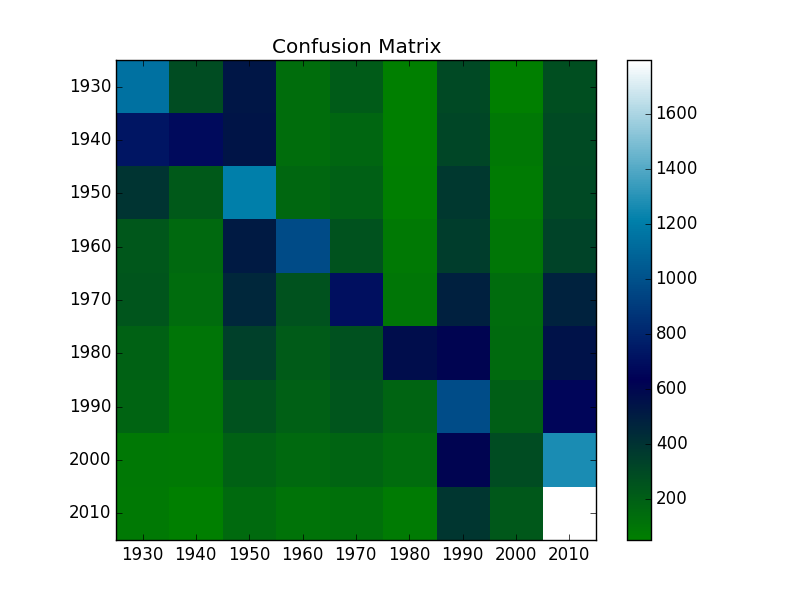
**2K)** Our classifier is 31% accurate. The expected accuracy of random chance is 1/9, or 11%. The expected accuracy of a uniformly random guess in the original balanced dataset could be calculated as P(Y), or the number of movies in the original unbalanced dataset from decade Y over the total number of movies in the original unbalanced dataset.

While our results are far from spectacular, they are an almost 3x improvement over random chance, indicating that our learning algorithm does provide some benefit. I suspect the accuracy is not higher because we implemented what you might call a *naïve naïve* Bayes Classifier (naïve squared!). Jokes aside, we reduced the granularity of our model by letting X­j­ = (word j appears in *i’s* plot summary), instead of X­j­ = (number of times word W­j­ appears in *i’*s plot summary). Namely, we only considered the probability that a word does or does not appear in a summary, not the number of times it appears in the summary. I suspect that had we done so, our accuracy would greatly improve.

**2L)** Please see below for the Cumulative Match Curve.



**2K)** Please see below for the confusion matrix. It seems that the most common mistake is confusing a 2000 movie with a 2010 movie.



**Problem 3**

**3a)** Please see below for the 10 most informative words for each decade. Weirdly, they seem to be very similar for each decade. This may explain the overall relatively low accuracy of our classifier—the most informative words are too similar between the various different decades.

1930: ('the', 'a', 'and', 'to', 'of', 'is', 'in', 'his', 'he', 'her')

1940: ('the', 'a', 'and', 'to', 'of', 'is', 'in', 'his', 'he', 'with')

1950: ('the', 'to', 'a', 'and', 'of', 'is', 'in', 'his', 'he', 'for')

1960: ('the', 'to', 'a', 'and', 'of', 'is', 'in', 'his', 'he', 'for')

1970: ('the', 'to', 'a', 'and', 'of', 'is', 'in', 'his', 'he', 'her')

1980: ('the', 'to', 'and', 'a', 'of', 'is', 'in', 'his', 'he', 'her')

1990: ('the', 'to', 'a', 'and', 'of', 'is', 'in', 'his', 'he', 'her')

2000: ('the', 'a', 'and', 'to', 'of', 'in', 'is', 'his', 'with', 'her')

2010: ('the', 'a', 'and', 'to', 'of', 'in', 'is', 'his', 'with', 'for')

**3b)** The accuracy decreased from 31% to 22%, a pretty substantial decrease.