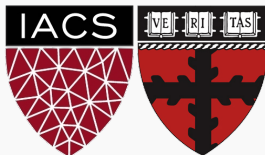


Lecture #11: Logistic Regression - Part II

Data Science 1

CS 109A, STAT 121A, AC 209A, E-109A

Pavlos Protopapas Kevin Rader
Margo Levine Rahul Dave



Logistic Regression: a Brief Review

Classification Boundaries

Regularization in Logistic Regression

Multinomial Logistic Regression

Bayes Theorem and Misclassification Rates

ROC Curves

Logistic Regression: a Brief Review



Multiple Logistic Regression

Earlier we saw the general form of *simple* logistic regression, meaning when there is just one predictor used in the model. What was the model statement (in terms of linear predictors)?

Multiple Logistic Regression

Earlier we saw the general form of *simple* logistic regression, meaning when there is just one predictor used in the model. What was the model statement (in terms of linear predictors)?



$$\log \left(\frac{P(Y = 1)}{1 - P(Y = 1)} \right) = \beta_0 + \beta_1 X$$

Multiple Logistic Regression

Earlier we saw the general form of *simple* logistic regression, meaning when there is just one predictor used in the model. What was the model statement (in terms of linear predictors)?

$$\log \left(\frac{P(Y = 1)}{1 - P(Y = 1)} \right) = \beta_0 + \beta_1 X$$

Multiple logistic regression is a generalization to multiple predictors. More specifically we can define a multiple logistic regression model to predict $P(Y = 1)$ as such:

$$\log \left(\frac{P(Y = 1)}{1 - P(Y = 1)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

where there are p predictors: $X = (X_1, X_2, \dots, X_p)$.

Note: statisticians are often lazy and use the notation \log to mean \ln (the text does this). We will write \log_{10} if this is what we mean.

Interpreting Multiple Logistic Regression: an Example

Let's get back to the NFL data. We are attempting to predict whether a play results in a TD based on location (yard line) and whether the play was a pass. The simultaneous effect of these two predictors can be brought into one model.

Recall from earlier we had the following estimated models:

$$\log \left(\frac{\widehat{P(Y = 1)}}{1 - \widehat{P(Y = 1)}} \right) = -7.425 + 0.0626 \cdot X_{yard}$$

$$\log \left(\frac{\widehat{P(Y = 1)}}{1 - \widehat{P(Y = 1)}} \right) = -4.061 + 1.106 \cdot X_{pass}$$

The results for the multiple logistic regression model are on the next slide.

Interpreting Multiple Logistic Regression: an Example




```
# Create data frame of predictors
X = nfldata[["YardLine", "IsPass"]]

# Create logistic regression object
logitm = sk.LogisticRegression(C = 1000000)
logitm.fit (X, nfldata["IsTouchdown"])

# The coefficients
print('Estimated beta1, beta2: \n', logitm.coef_)
print('Estimated beta0: \n', logitm.intercept_)
```

```
Estimated beta1, beta2:
[[ 0.06547811  1.2066147 ]]
Estimated beta0:
[-8.30059191]
```


Some questions

1. Write down the complete model. Break this down into the model to predict log-odds of a touchdown based on the yard line for passes and the same model for non-passes. How is this different from the previous model (without interaction)? 
2. Estimate the odds ratio of a TD comparing passes to non-passes.
3. Is there any evidence of multicollinearity in this model? 
4. Is there any confounding in this problem? 

Interactions in Multiple Logistic Regression

Just like in linear regression, interaction terms can be considered in logistic regression.

An interaction terms is incorporated into the model the same way, and the interpretation is very similar (on the log-odds scale of the response of course).



Write down the model for the NFL data for the 2 predictors plus the interactions term.

Interpreting Multiple Logistic Regression with Interaction: an Ex

```
# Create data frame of predictors
nfldata['Interaction'] = nfldata["YardLine"]*nfldata["IsPass"]
X = nfldata[["YardLine", "IsPass", "Interaction"]]

# Create logistic regression object
logitm = sk.LogisticRegression(C = 10000000000000000)
logitm.fit(X, nfldata["IsTouchdown"])

# The coefficients
print('Estimated beta1, beta2, beta3: \n', logitm.coef_)
print('Estimated beta0: \n', logitm.intercept_)

nfldata['Intercept'] = 1.0
logit_sm = sm.Logit(nfldata['IsTouchdown'], nfldata[["Intercept",
YardLine", "IsPass", "Interaction"]])
fit_sm = logit_sm.fit()
print(fit_sm.summary())

nfldata.head()
```


	YardLine	IsPass	Interaction
46259	57	0	0
3778	73	1	73
20707	34	0	0
45826	83	1	83
10982	78	1	78

Estimated beta1:
[[0.06769992 1.46499967 -0.00319916]]

Estimated beta0:
[-8.48339235]



Some questions

1. Write down the complete model. Break this down into the model to predict log-odds of a touchdown based on the yard line for passes and the same model for non-passes. How is this different from the previous model (without interaction)? 
2. Use this model to estimate the probability of a touchdown for a pass at the 20 yard line. Do the same for a run at the 20 yard line.
3. Use this model to estimate the probability of a touchdown for a pass at the 99 yard line. Do the same for a run at the 99 yard line.
4. Is this a stronger model than the previous one? How would we check?

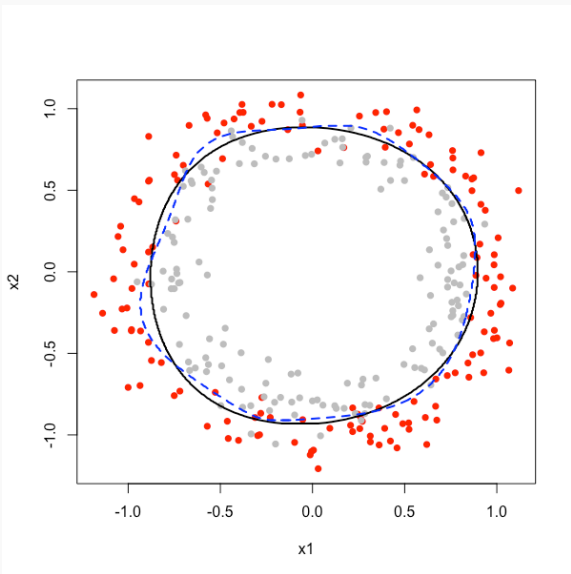
Classification Boundaries

Recall that we could attempt to purely classify each observation based on whether the estimated $P(Y = 1)$ from the model was greater than 0.5.

When dealing with ‘well-separated’ data, logistic regression can work well in performing classification.

We saw a 2-D plot last time which had two predictors, X_1 and X_2 and depicted the classes as different colors. A similar one is shown on the next slide.

2D Classification in Logistic Regression: an Example



2D Classification in Logistic Regression: an Example


Would a logistic regression model perform well in classifying the observations in this example?

2D Classification in Logistic Regression: an Example

Would a logistic regression model perform well in classifying the observations in this example?

What would be a good logistic regression model to classify these points?

2D Classification in Logistic Regression: an Example

Would a logistic regression model perform well in classifying the observations in this example? 

What would be a good logistic regression model to classify these points?

Based on these predictors, two separate logistic regression model were considered that were based on different ordered polynomials of X_1 and X_2 and their interactions. The 'circles' represent the boundary for classification.



How can the classification boundary be calculated for a logistic regression?

2D Classification in Logistic Regression: an Example

In the previous plot, which classification boundary performs better? How can you tell? How would you make this determination in an actual data example?

2D Classification in Logistic Regression: an Example

In the previous plot, which classification boundary performs better? How can you tell? How would you make this determination in an actual data example?

We could determine the misclassification rates in left out validation or test set(s)

Regularization in Logistic Regression

Regularization in Linear Regression

Based on the Likelihood framework, a loss function can be determined based on the likelihood function.

We saw in linear regression that maximizing the log-likelihood is equivalent to minimizing the sum of squares error:

Regularization in Linear Regression

Based on the Likelihood framework, a loss function can be determined based on the likelihood function.

We saw in linear regression that maximizing the log-likelihood is equivalent to minimizing the sum of squares error:

$$\arg \min \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \arg \min \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}))^2$$

And a regularization approach was to add a penalty factor to this equation. Which for Ridge Regression becomes:

Regularization in Linear Regression

Based on the Likelihood framework, a loss function can be determined based on the likelihood function.

We saw in linear regression that maximizing the log-likelihood is equivalent to minimizing the sum of squares error:

$$\arg \min \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \arg \min \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}))^2$$

And a regularization approach was to add a penalty factor to this equation. Which for Ridge Regression becomes:

$$\arg \min \left[\sum_{i=1}^n \left(y_i - \left(\beta_0 + \sum_{j=1}^n \beta_j x_{ji} \right) \right)^2 + \lambda \sum_{j=1}^n \beta_j^2 \right]$$

This penalty *shrinks* the estimates towards zero, and had the analogue of using a Normal prior in the Bayesian paradigm.

Loss function in Logistic Regression

A similar approach can be used in logistic regression. Here, maximizing the log-likelihood is equivalent to minimizing the following loss function:

$$\arg \min \left[- \sum_{i=1}^n (y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)) \right]$$

where $\hat{p}_i = \frac{\exp(\beta_0 + \sum_{j=1}^n \beta_j x_{ji})}{1 + \exp(\beta_0 + \sum_{j=1}^n \beta_j x_{ji})}$.

Why is this a good loss function to minimize? Where does this come from?

Loss function in Logistic Regression

A similar approach can be used in logistic regression. Here, maximizing the log-likelihood is equivalent to minimizing the following loss function:

$$\arg \min \left[- \sum_{i=1}^n (y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)) \right]$$

where $\hat{p}_i = \frac{\exp(\beta_0 + \sum_{j=1}^n \beta_j x_{ji})}{1 + \exp(\beta_0 + \sum_{j=1}^n \beta_j x_{ji})}$.

Why is this a good loss function to minimize? Where does this come from?

The log-likelihood for independent $Y_i \sim \text{Bern}(p_i)$:

Regularization in Logistic Regression

A penalty factor can then be added to this loss function and results in a new loss function that penalizes large values of the parameters:

$$\arg \min \left[- \sum_{i=1}^n [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)] + \lambda \sum_{j=1}^n \beta_j^2 \right]$$

The result is just like in linear regression: shrinkage towards zero of the parameters.

In practice, the intercept is usually not part of the penalty factor, and is thus not shrunk towards zero.

Note: the `sklearn` package uses a different tuning parameter: instead of λ they use a constant that is essentially $C = 1/\lambda$.

Regularization in Logistic Regression: an Example

Let's see how this plays out in an example in logistic regression.

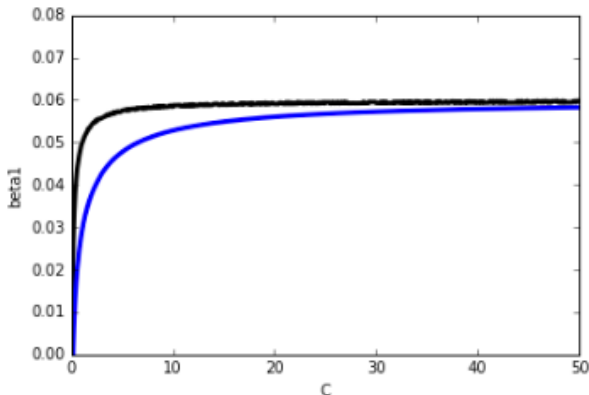
Regularization in Logistic Regression: an Example

```
betal_11 = []
betal_12 = []
Cs = []
X = polynomial_basis (nfldata_sm["YardLine"], 1)

for i in range(1, 500):
    C = i/10
    logitm_11 = sk.LogisticRegression(C = C, penalty = "l1")
    logitm_11.fit (X, nfldata_sm["IsTouchdown"])
    logitm_12 = sk.LogisticRegression(C = C, penalty = "l2")
    logitm_12.fit (X, nfldata_sm["IsTouchdown"])
    betal_11.append(logitm_11.coef_[0])
    betal_12.append(logitm_12.coef_[0])
    Cs.append(C)
```

Regularization in Logistic Regression: an Example

```
plt.plot(Cs, betal_11, color='black', lw=3)  
plt.plot(Cs, betal_12, color='blue', lw=3)  
plt.xlabel("C")  
plt.ylabel("betal")  
plt.ylim(0,0.08)  
plt.show()
```



Regularization in Logistic Regression: an Example

Just like in linear regression, the shrinkage factor must be chosen. How should we go about doing this?

Regularization in Logistic Regression: an Example

Just like in linear regression, the shrinkage factor must be chosen. How should we go about doing this?



Through building multiple training and test sets (through k -fold or random subsets), we can select the best shrinkage factor to mimic out-of-sample prediction.

How could we measure how well each model fits the test set?
We could measure this based on the proposed loss function!

Multinomial Logistic Regression

Logistic Regression for predicting more than 2 Classes

There are several extensions to standard logistic regression when the response variable Y has more than 2 categories.

The two most common are :

1. ordinal logistic regression
2. multinomial logistic regression.

Logistic Regression for predicting more than 2 Classes

There are several extensions to standard logistic regression when the response variable Y has more than 2 categories.

The two most common are :

1. ordinal logistic regression
2. multinomial logistic regression.

Ordinal logistic regression is used when the categories have a specific hierarchy (like class year: Freshman, Sophomore, Junior, Senior; or a 7-point rating scale from strongly disagree to strongly agree).

Multinomial logistic regression is used when the categories have no inherent order (like eye color: blue, green, brown, hazel, et...).



Multinomial Logistic Regression

The most common approach to estimating a nominal (not-ordinal) categorical variable that has more than 2 classes. The first approach sets one of the categories in the response variable as the *reference* group, and then fits separate logistic regression models to predict the other cases based off of the reference group. For example we could attempt to predict a student's concentration:

$$y = \begin{cases} 1 & \text{if Computer Science (CS)} \\ 2 & \text{if Statistics} \\ 3 & \text{otherwise} \end{cases} .$$

from predictors x_1 number of psets per week and x_2 how much time spent in Lamont Library.

Multinomial Logistic Regression (cont.)

We could select the $y = 3$ case as the reference group (other concentration), and then fit two separate models: a model to predict $y = 1$ (CS) from $y = 3$ (others) and a separate model to predict $y = 2$ (Stat) from $y = 3$ (others).

Ignoring interactions, how many parameters would need to be estimated?

How could these models be used to estimate the probability of an individual falling in each concentration?

One vs. Rest (ovr) Logistic Regression (cont.)

The default multiclass logistic regression model is called the 'One vs. Rest' approach.

If there are 3 classes, then 3 separate logistic regressions are fit, where the probability of each category is predicted over the rest of the categories combined. So for the concentration example, 3 models would be fit:

1. a first model would be fit to predict CS from (Stat and Others) combined
2. a second model would be fit to predict Stat from (CS and Others) combined
3. a third model would be fit to predict Others from (CS and Stat) combined

An example to predict play call from the NFL data follows...

OVR Logistic Regression in Python

```
X = polynomial_basis (nfldata["YardLine"], 1)

nfldata["PlayType"] = nfldata["IsPass"] + 2 * nfldata["IsRush"]

logitm = sk.LogisticRegression(C = 10000000)
logitm.fit (X, nfldata["PlayType"])

# The coefficients
print('Estimated beta1: \n', logitm.coef_)
print('Estimated beta0: \n', logitm.intercept_)
```

```
Estimated beta1:
[[-0.01460736]
 [ 0.00635893]
 [ 0.00652455]]
Estimated beta0:
[-0.26422696 -0.61186328 -1.20051275]
```

Classification for more than 2 Categories

When there are more than 2 categories in the response variable, then there is no guarantee that $P(Y = k) \geq 0.5$ for any one category. So any classifier based on logistic regression will instead have to select the group with the largest estimated probability.

The classification boundaries are then much more difficult to determine. We will not get into the algorithm for drawing these in this class.

Bayes Theorem and Misclassification Rates

We defined conditional probability as:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

And using the fact that $P(B \cap A) = P(A|B)P(B)$ we get Bayes' Theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Another version of Bayes' Theorem is found by substituting in the Law of Total Probability (LOTP) into the denominator:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^C)P(B^C)}$$



Where have we seen Bayes' Theorem before? Why do we care?

In the diagnostic testing paradigm, one cares about whether the results of a test (like a classification test) matches truth (the true class that observation belongs to). The simplest version of this is trying to detect disease ($D+$ vs. $D-$) based on a diagnostic test ($T+$ vs. $T-$).

Medical examples of this include various screening tests: breast cancer screening through (i) self-examination and (ii) mammographies, prostate cancer screening through (iii) PSA tests, and Colo-rectal cancer through (iv) colonoscopies.

These tests are a little controversial because of poor predictive probability of the tests.

Bayes' theorem can be rewritten for diagnostic tests:

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)}$$

These probability quantities can then be defined as:

- ▶ **Sensitivity:** $P(T+|D+)$
- ▶ **Specificity:** $P(T-|D-)$
- ▶ **Prevalence:** $P(D+)$
- ▶ **Positive Predictive Value:** $P(D+|T+)$
- ▶ **Negative Predictive Value:** $P(D-|T-)$

How do positive and negative predictive values relate? Be careful...

Diagnostic Testing (cont.)

We mentioned that these tests are a little controversial because of their poor predictive probability. When will these tests have poor positive predictive probability?

Diagnostic Testing (cont.)

We mentioned that these tests are a little controversial because of their poor predictive probability. When will these tests have poor positive predictive probability?

When the disease is not very prevalent, then the number of 'false positives' will overwhelm the number of true positive. For example, PSA screening for prostate cancer has sensitivity of about 90% and specificity of about 97% for some age groups (men in their fifties), but prevalence is about 0.1%.

What is positive predictive probability for this diagnostic test?

Why do we care?

As data scientists, why do we care about diagnostic testing from the medical world? (hint: it's not just because Kevin is a trained biostatistician!)

Why do we care?

As data scientists, why do we care about diagnostic testing from the medical world? (hint: it's not just because Kevin is a trained biostatistician!)

Because classification can be thought of as a diagnostic test.

Let $Y_i = k$ be the event that observation i truly belongs to category k , and let $\hat{Y}_i = k$ be the event that we correctly predict it to be in class k . Then Bayes' rule states that our *Positive Predictive Value* for classification is:

$$P(Y_i = k | \hat{Y}_i = k) = \frac{P(\hat{Y}_i = k | Y_i = k)P(Y_i = k)}{P(\hat{Y}_i = k | Y_i = k)P(Y_i = k) + P(\hat{Y}_i = k | Y_i \neq k)P(Y_i \neq k)}$$

Thus the probability of a predicted outcome truly being in a specific group depends on what?

Why do we care?

As data scientists, why do we care about diagnostic testing from the medical world? (hint: it's not just because Kevin is a trained biostatistician!)

Because **classification can be thought of as a diagnostic test.**

Let $Y_i = k$ be the event that observation i truly belongs to category k , and let $\hat{Y}_i = k$ be the event that we correctly predict it to be in class k . Then Bayes' rule states that our *Positive Predictive Value* for classification is:

$$P(Y_i = k | \hat{Y}_i = k) = \frac{P(\hat{Y}_i = k | Y_i = k)P(Y_i = k)}{P(\hat{Y}_i = k | Y_i = k)P(Y_i = k) + P(\hat{Y}_i = k | Y_i \neq k)P(Y_i \neq k)}$$

Thus the probability of a predicted outcome truly being in a specific group depends on what? The proportion of observations in that class!

There are 2 major types of error in classification problems based on a binary outcome. They are:

- ▶ False positives: incorrectly predicting $\hat{Y} = 1$ when it truly is in $Y = 0$.
- ▶ False negative: incorrectly predicting $\hat{Y} = 0$ when it truly is in $Y = 1$.

The results of a classification algorithm are often summarized in two ways: a confusion table, sometimes called a contingency table, or a 2x2 table (more generally $k \times k$ table) and an receiver operating characteristics (ROC) curve.

Confusion table

When a classification algorithm (like logistic regression) is used, the results can be summarize in a $k \times k$ table as such:

		True Republican Status	
		Yes	No
Predicted	Yes	487	288
Republican	No	218	314



The table above was a classification based on a logistic regression model to predict political party (Dem. vs. Rep.) based on 3 predictors: X_1 = whether respondent believes abortion is legal, X_2 = income (logged) and X_3 = years of education.

What are the false positive and false negative rates for this classifier?

Bayes' Classifier Choice

A classifier's error rates can be tuned to modify this table.
How?

Bayes' Classifier Choice

A classifier's error rates can be tuned to modify this table.
How?

The choice of the Bayes' classifier level will modify the characteristics of this table.

If we thought it was more important to predict republicans correctly (lower false positive rate), what could we do for our Bayes' classifier level?

Bayes' Classifier Choice

A classifier's error rates can be tuned to modify this table.
How?

The choice of the Bayes' classifier level will modify the characteristics of this table.

If we thought it was more important to predict republicans correctly (lower false positive rate), what could we do for our Bayes' classifier level?



We could classify instead based on:

$$\hat{P}(Y = 1) < \pi$$

and we could choose π to be some level other than 0.5. Let's see what the table looks like if π were 0.28 or 0.52 instead (why such strange numbers?).

Other Confusion table

Based on $\pi = 0.28$:

		True Republican Status	
		Yes	No
Predicted	Yes	247	528
Republican	No	80	452

What has improved? What has worsened?

Other Confusion table

Based on $\pi = 0.28$:

		True Republican Status	
		Yes	No
Predicted	Yes	247	528
Republican	No	80	452



What has improved? What has worsened?

Based on $\pi = 0.52$:

		True Republican Status	
		Yes	No
Predicted	Yes	627	148
Republican	No	388	144



Which should we choose? Why?

ROC Curves

The ROC curve illustrates the trade-off for all possible thresholds chosen for the two types of error (or correct classification).

The vertical axis displays the true positive predictive value and the horizontal axis depicts the true negative predictive value.

What is the shape of an ideal ROC curve?

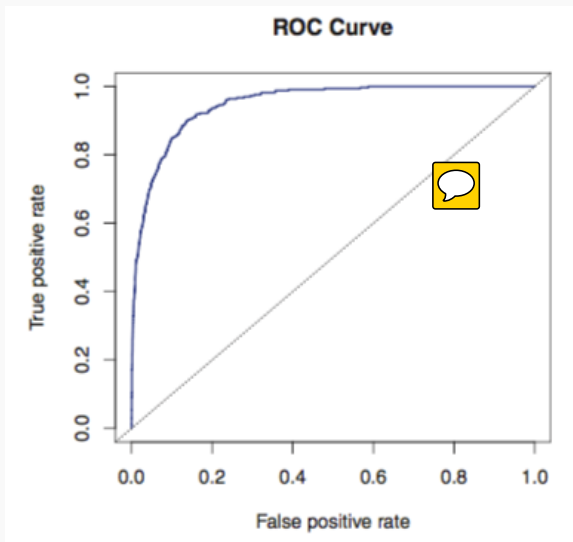
The ROC curve illustrates the trade-off for all possible thresholds chosen for the two types of error (or correct classification).

The vertical axis displays the true positive predictive value and the horizontal axis depicts the true negative predictive value.

What is the shape of an ideal ROC curve?

See next slide for an example.

ROC Curve Example



ROC Curve for measuring classifier performance

The overall performance of a classifier, calculated over all possible thresholds, is given by the area under the ROC curve ('AUC').

An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier.

What is the worst case scenario for AUC? What is the best case? What is AUC if we independently just flip a coin to perform classification?

ROC Curve for measuring classifier performance

The overall performance of a classifier, calculated over all possible thresholds, is given by the area under the ROC curve ('AUC').

An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier.

What is the worst case scenario for AUC? What is the best case? What is AUC if we independently just flip a coin to perform classification?

This AUC then can be used to compare various approaches to classification: Logistic regression, LDA (to come), kNN, etc...