Your name: Ross Alexander

Your SUNet ID: rbalexan@stanford.edu (06353460)

Exam rules:

- You have until 4:00 PM July 22, 2021 to complete the exam and submit it to Gradescope.
- Following the updated BJA guidelines for open book exams, you are allowed to use your textbook, course slides, notes, and the internet in completing this exam.
- A Cheat Sheet is provided at the end of the exam.
- Please show your work and justify your answers.

Problem	Points	Max
1		20
2		10
3		10
4		10
5		20
6		10
7		20
Total		100

1. We define a new kind of discriminant analysis for a classification problem with a binary response. The classes have prior probabilities π_0 and π_1. Given the class, k, the conditional probability of the inputs $X_1, \ldots, X_p$ is multivariate normal with a class-dependent mean μ_k and covariance matrix $\sigma_k \mathbf{\Sigma}$. The matrix $\mathbf{\Sigma}$ is common to both classes and σ_k is a class-dependent constant. All parameters, π_k, μ_k, σ_k, for each class, as well as Σ, are set to their Maximum Likelihood estimates.

   (a) **[10 points]** Provide an equation describing the classifier's decision boundary or discriminant. What would the boundary look like?

Leveraging the log-posterior of class k given an input x in QDA and modifying the values to suit our problem, we have

$\mathcal{L}_0$: (class 0): $C + \log \pi_0 - \frac{1}{2}\mu_0^T(\sigma_0\Sigma)^{-1}\mu_0 + x^T(\sigma_0\Sigma)^{-1}\mu_0 - \frac{1}{2}x^T(\sigma_0\Sigma)^{-1}x - \frac{1}{2}\log|\sigma_0\Sigma|$

$\mathcal{L}_1$: (class 1): $C + \log \pi_1 - \frac{1}{2}\mu_1^T(\sigma_1\Sigma)^{-1}\mu_1 + x^T(\sigma_1\Sigma)^{-1}\mu_1 - \frac{1}{2}x^T(\sigma_1\Sigma)^{-1}x - \frac{1}{2}\log|\sigma_1\Sigma|$

The decision boundary is the locus of all points where the posterior probability (or log-posterior) of the two classes is equal, i.e. $\mathcal{L}_0 = \mathcal{L}_1$

$$\cancel{C} + \log\pi_0 - \frac{1}{2\sigma_0}\mu_0^T\Sigma^{-1}\mu_0 + \frac{1}{\sigma_0}x^T\Sigma^{-1}\mu_0 - \frac{1}{\sigma_0}x^T\Sigma^{-1}x - \frac{1}{2}\log(\sigma_0^n|\Sigma|) =$$
$$= \cancel{C} + \log\pi_1 - \frac{1}{2\sigma_1}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{\sigma_1}x^T\Sigma^{-1}\mu_1 - \frac{1}{\sigma_1}x^T\Sigma^{-1}x - \frac{1}{2}\log(\sigma_1^n|\Sigma|)$$
$$0 = \log\frac{\pi_0}{\pi_1} - \frac{1}{2\sigma_0}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2\sigma_1}\mu_1^T\Sigma^{-1}\mu_1 + x^T\Sigma^{-1}\left(\frac{\mu_0}{\sigma_0} - \frac{\mu_1}{\sigma_1}\right) - \left(\frac{1}{\sigma_0} - \frac{1}{\sigma_1}\right)x^T\Sigma^{-1}x - \frac{1}{2}\log\left(\frac{\sigma_0^n|\Sigma|}{\sigma_1^n|\Sigma|}\right)$$

(first group: let equal $\tilde{c}$)

$$0 = -\left(\frac{1}{\sigma_0} - \frac{1}{\sigma_1}\right)x^T\Sigma^{-1}x + \left(\frac{\mu_0}{\sigma_0} - \frac{\mu_1}{\sigma_1}\right)^T\Sigma^{-1}x + \tilde{c}$$

quadratic (paraboloid) if $\sigma_0 \neq \sigma_1$, linear (hyperplane) if $\sigma_0 = \sigma_1$

   (b) **[5 points]** Why might this classifier be preferable to Linear Discriminant Analysis?

> This classifier might be preferable to LDA since it has additional model flexibility – we can allow the two classes to have differently-scaled versions of a covariance matrix rather than forcing the two classes to have identically-scaled (identical) covariance matrices.

   (c) **[5 points]** Why might this classifier be preferable to Quadratic Discriminant Analysis?

> This classifier might be preferable to QDA since it has fewer parameters/DoFs to manage. In situations with little data or if we are in search of a more parsimonious model, this classifier has fewer parameters than QDA an will be easier to fit and explain. It is also possible that we might have domain knowledge that the classes have a similar covariance structure even if they have different scales. In this case, we would prefer our classifier over QDA to enforce the domain knowledge constraint.

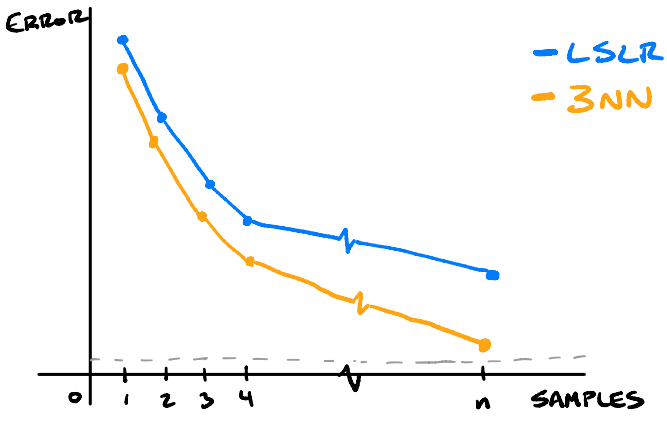
2. [**10 points**] Compare leave-one-out cross validation to 10-fold cross validation, with reference to the bias-variance tradeoff.

LOOCV will fit models with low variance but high bias since all but one of the points is used to fit the model each time. As a result, the model fit is very similar each time (low variance) but biased towards the data in the training set. 10-fold CV will fit models with larger variance and smaller bias than LOOCV since there are larger portions of the dataset that are withheld in each fit and the training set is more different between each CV iteration than in LOOCV. In estimating the model error, we expect LOOCV to be low bias, but high variance due to correlation of the model fits, whereas in 10-fold CV, we expect higher bias, but lower variance due to how the dataset is partitioned into folds.

3. [**10 points**] A total of n samples were simulated from the following distribution

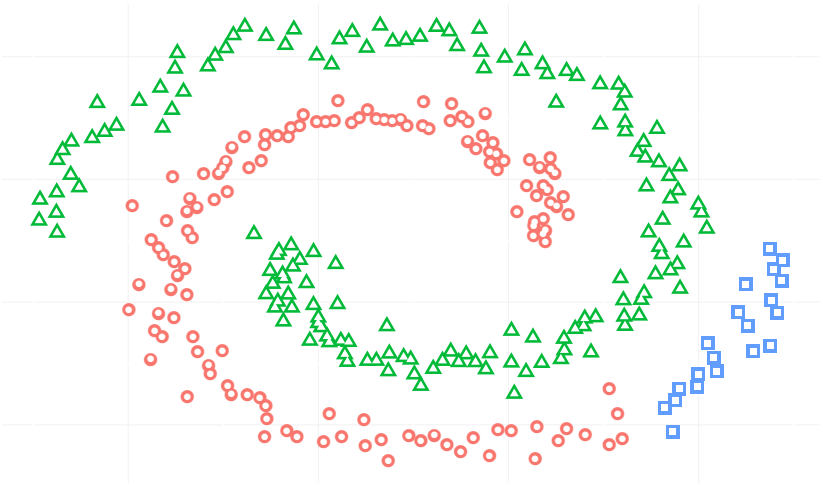
$$X_1, X_2, X_3, X_4 \sim \mathcal{N}(0, 1) \text{ i.i.d.}$$
$$Y = X_1 + 2X_2 + X_3^3 + X_1X_4 + \epsilon,$$

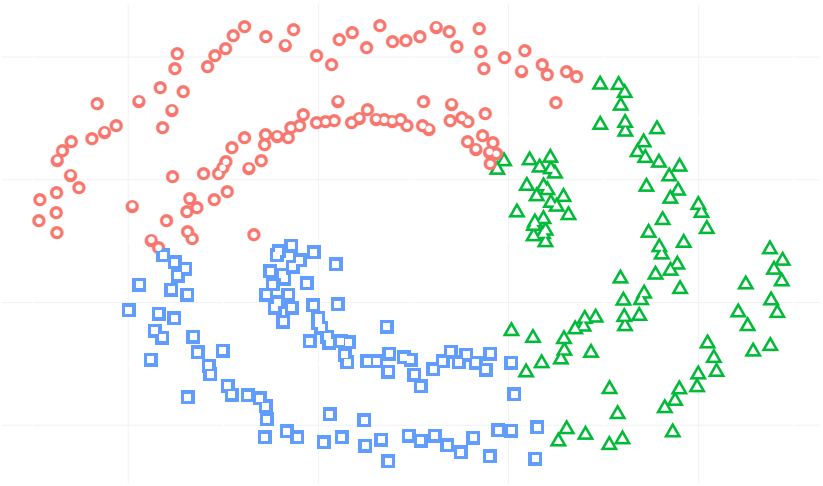
where f is non-linear. Consider the following regression methods for Y: linear regression with predictors X_1, X_2, X_3, and X_4, and 3-nearest neighbors regression. On the same plot, sketch a plausible learning curve for each method. A learning curve for regression shows the average test MSE as a function of n. Explain your reasoning.



We have very little information about other aspects of the problem, so the sketches are very general. For both methods, as we increase the number of samples, our model's test error decreases. The least-squares linear regression (LSLR) model is not consistent with the underlying model, so while the linear fit will get better with more samples, it will have large test error even at n samples. The 3-nearest neighbors (3NN) model will get better as the samples start to fill the feature space and there are more local examples to draw on as neighbors. After n samples, we might expect 3NN to have lower test error than the LSLR model, since 3NN will be more flexible (as it is a nonparametric model). However, if the sample variance is large, 3NN may not work as great as LSLR. As a general thought though, we expect 3NN to perform better than LSLR at n samples.

4. **[10 points]** The clusterings below were produced by single-linkage hierarchical clustering and k-means clustering. Determine which one is which and explain your reasoning.





The upper figure shows a clustering produced by single-linkage hierarchical clustering and the lower figure shows a clustering produced by k-means clustering. It is easy to differentiate the two methods in this case, since k-means clustering relies on cluster centroids and a distance metric. For k-means clustering, the points closest to a particular cluster centroid will be classified as that particular cluster. We can see in the upper figure that this property is not observed – for example, the intertwined spiral structure of the red and green clusters would not be possible to achieve using k-means clustering because the clusters cannot "overlap" in the feature space.

5. We apply the Bootstrap to a dataset with n distinct observations $x_1, \ldots, x_n$.

   (a) **[10 points]** What is the probability that the j^{th} observation is included in a specific bootstrap sample?

The probability that the jth observation is not the first sample in a bootstrap sample is

$$P(x_j \neq X_i^b) = 1 - \frac{1}{n}$$

Since the bootstrap method samples observations independently with replacement, then the probability the jth observation is not in the bootstrap sample is

$$P(x_j \notin X^b) = \prod_{i=1}^{n} P(x_j \neq X_i^b)$$
$$= \left(1 - \frac{1}{n}\right)^n$$

Therefore, the probability the jth observation is in the bootstrap sample is

$$P(x_j \in X^b) = 1 - P(x_j \notin X^b)$$
$$= 1 - \left(1 - \frac{1}{n}\right)^n$$

   (b) **[10 points]** What is the expected value for the fraction of distinct observations in a specific bootstrap sample (i.e. the number of distinct observations from $x_1, \ldots, x_n$ divided by n). Does this expectation converge as n grows large? Hints: (i) use the probability from part (a), (ii) $\lim_{n\to\infty}(1 - 1/n)^n = e^{-1}$.

The expected value for the fraction of distinct observations in a specific bootstrap sample is

$$\frac{\sum_{i=1}^{n} \mathbb{E}[\mathbb{1}\{x_i \in X^b\}]}{n}$$

Since the probability of observations being included in the bootstrap sample is independent of the observations, we have

$$\frac{\sum_{i=1}^{n} \mathbb{E}[\mathbb{1}\{x_i \in X^b\}]}{n} = \frac{n\,\mathbb{E}[\mathbb{1}\{x_i \in X^b\}]}{n} \quad \text{for any } i \in 1, 2, \ldots, n$$
$$= \mathbb{E}[\mathbb{1}\{x_i \in X^b\}]$$
$$= (0)P(x_i \notin X^b) + (1)P(x_i \in X^b)$$
$$= P(x_i \in X^b)$$
$$= 1 - \left(1 - \frac{1}{n}\right)^n$$

In the limit of $n \to \infty$

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mathbb{E}[\mathbb{1}\{x_i \in X^b\}]}{n} = \lim_{n\to\infty} 1 - \left(1 - \frac{1}{n}\right)^n$$
$$= 1 - \lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n$$
$$\overset{5}{=} 1 - \frac{1}{e}$$

which converges in the limit of large n.

6. [**10 points**] A group of 33 people were asked to report their happiness on a scale from 0 to 20. We apply a linear model with an intercept to regress happiness onto 2 predictors, the yearly income and the amount of money paid in taxes last year.

   The t-statistics for income and taxes have corresponding p-values of 0.14 and 0.52, respectively. The RSS of the model is 30 and the sample variance of the happiness is 1.5.

   What would you conclude about the relationship between happiness, income, and tax contributions?

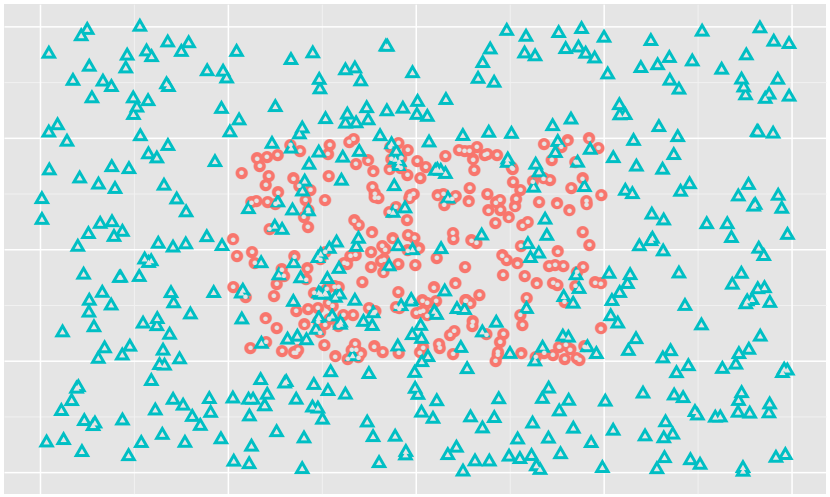
   Looking at the sample variance of 1.5, we can see that the average squared deviation from the mean happiness is 1.5 and that the average absolute deviation from the mean happiness is approximately 1.25 (sqrt(1.5) ~= 1.25). This indicates that our happiness values are quite closely clustered, i.e. there is a relatively tight distribution of happiness values. Further, this means that there is probably relatively weak correspondence with the predictors (i.e. near-zero coefficients).

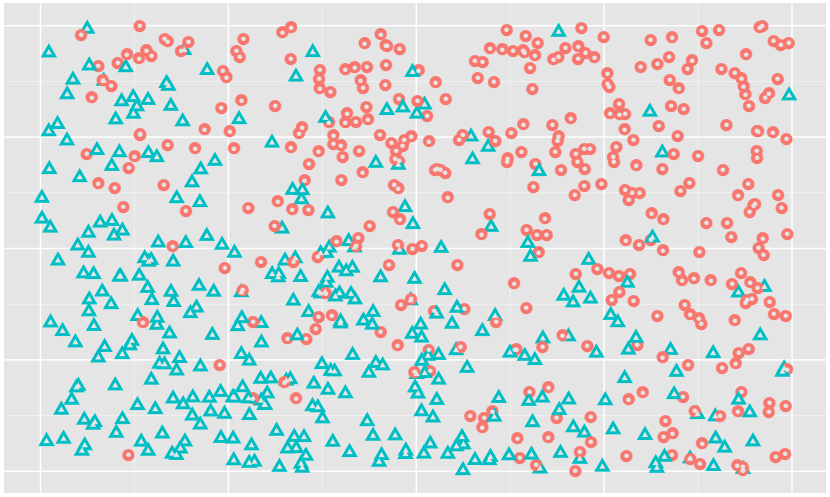
   We obtain a relatively low RSS. For the RSS of 30 on 33 observations, this implies an average squared residual of approximately 1 (30/33 ~= 1) and therefore an average absolute residual of approximately 1 (sqrt(1) = 1). Therefore, we are on average only missing the prediction by 1 happiness point, which at first seems quite good on the scale of 0-20. However, in the context of the sample variance, the average absolute deviation is approximately 1.25, so we are only doing slightly better than the irreducible noise of the dataset.

   Finally, examining the p-values, we see that under the standard definition ($p < 0.05$), the two predictors are not statistically significant, though we are close to being able to reject the null hypothesis for the income predictor. So under this model, we conclude there is not a statistically meaningful relationship between the predictors (income, tax contributions) with the target (happiness).

   Using domain knowledge, it is likely that the predictors are strongly correlated (collinear), so we might expect a regression of happiness on income alone or a regression of happiness on tax contributions alone to demonstrate that there is a statistically significant relationship between the predictor and the target.

7. (a) **[10 points]** Identify which classifier among k-nearest neighbors with $k = 15$ and logistic regression would be more appropriate for each dataset below. Explain how one might adjust the True Positive rate of each method.



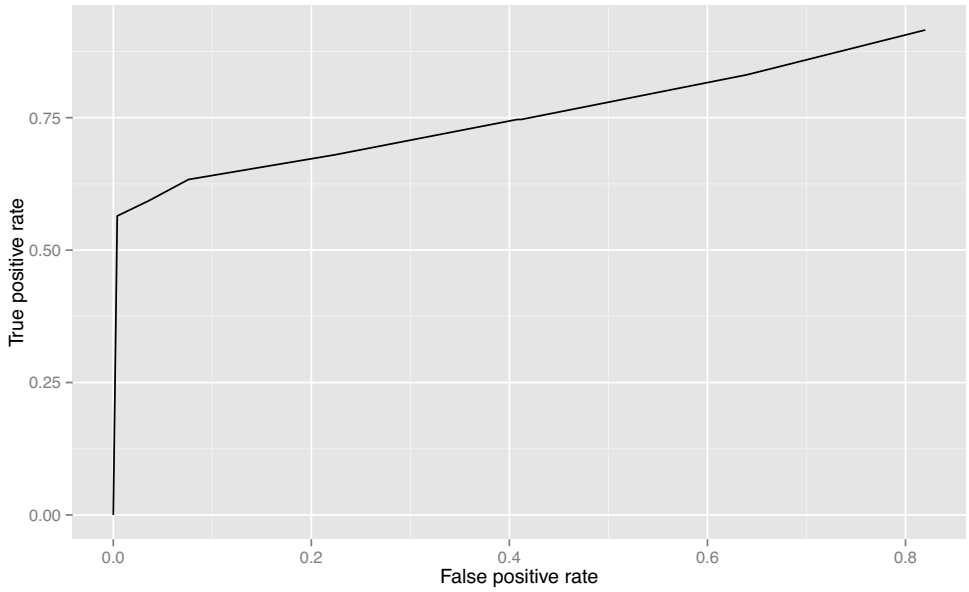


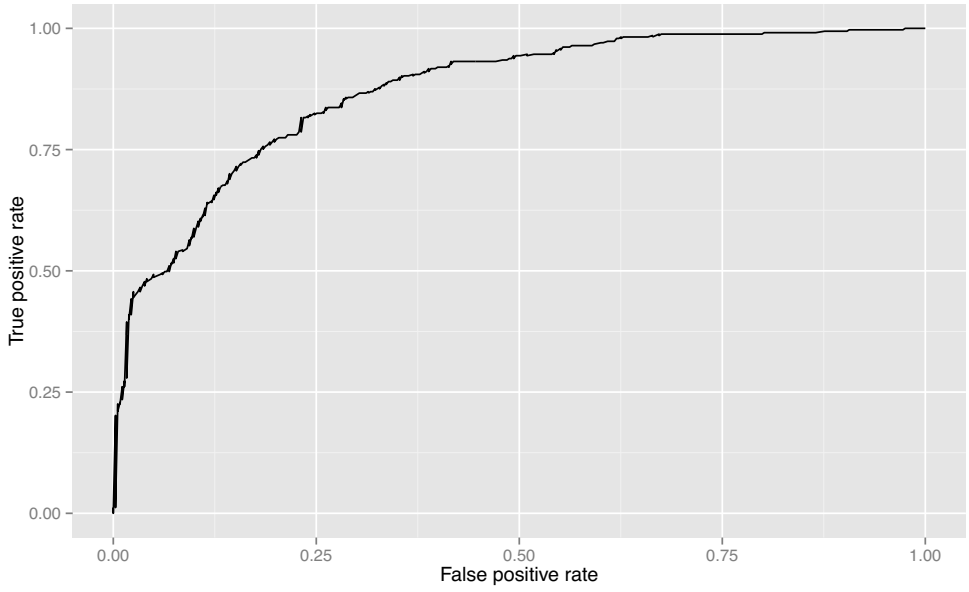
Note: Red circles are negative and blue triangles are positive.

The upper dataset should be classified using k-nearest neighbors (KNN) and the lower dataset should be classified using logistic regression. KNN works best when there is strong local structure since KNN is nonparametric. We see this structure in the upper dataset where there is an inner rectangle of negative examples and a larger outer rectangle of positive examples. Logistic regression produces a linear decision boundary and cannot respond to variations in local structure. It is easy to visualize a simple linear decision boundary on the lower dataset that would result in a good classifier.

For KNN with k=15, we can adjust the true positive rate by deciding the "majority votes" threshold. Instead of 8 positive examples being the threshold for classifying the test point as positive, we could set the threshold to more or less — 5 or 14, for example. (We could also adjust k, but I think that's not what you're asking). For logistic regression, we can adjust the true positive rate by adjusting the probability threshold for classification. While the threshold for classifying a test point as a positive example is typically 0.5, we could adjust it higher or lower – to 0.1, or 0.85, for example.

(b) **[10 points]** Each of the ROC curves below corresponds to one of the datasets in part (a). In each case, we applied the optimal classifier among k-nearest neighbors and logistic regression. Match each ROC curve to its corresponding dataset and explain your reasoning.





The upper ROC curve is for the upper dataset using a KNN classifier. Since the inner rectangle of negative examples is much denser than the positive examples in the same region, changing the "majority vote" threshold will only have sizeable effects at low thresholds when the number of positive examples finally exceeds the "majority vote" threshold. Moreover, there are only 16 possible choices for the threshold parameter, so the ROC curve can have at most 16 distinct portions (the lower plot clearly has many more).

The lower ROC curve is for the lower dataset using a logistic regression classifier. We can see that adjusting our probability threshold higher or lower (equivalent to shifting the decision boundary perpendicular to itself) results in better classification of one class of examples and worse classification of the other class of examples. We also rationalize achieving 0% TPR and 0% FPR if we trivially classify all examples as negative (decision boundary shifts to infinity at bottom-left of plot) and 100% TPR and 100% FPR if we trivially classify all examples as positive (decision boundary shifts to infinity at top-right of plot). Since there are infinite choices of the threshold (on interval [0, 1]), the number of distinct segments on the ROC curve is limited by the number of training examples. We have a lot of training examples and observe a lot of segments on the ROC curve, so we believe this is correct for the linear regression classifier on the lower dataset.

Cheat sheet

The sample variance of $x_1, \ldots, x_n$ is:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

The residual sum of squares for a regression model is:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

t-test:

The t-statistic for hypothesis $H_0 : \beta_i = 0$ is

$$t = \frac{\hat{\beta}_i}{\mathrm{SE}(\hat{\beta}_i)}$$

F-test:

The F-statistic for hypothesis $H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$ is

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)},$$

where RSS_0 is the residual sum of squares for the null model H_0, and RSS is the residual sum of squares for the full model with all predictors. Asymptotically, the F-statistic has the F-distribution with degrees of freedom $d_1 = q$ and $d_2 = n - p - 1$.

Minimum F-statistic to reject H_0 at a significance level $\alpha = 0.01$

		d_1			
		1	2	3	4
	1	4052.181	4999.500	5403.352	5624.583
	10	10.044	7.559	6.552	5.994
d_2	20	8.096	5.849	4.938	4.431
	30	7.562	5.390	4.510	4.018
	120	6.851	4.787	3.949	3.480

Logistic regression:

Logistic regression assigns to positive if the estimated conditional probability

$$\hat{P}(Y = +|X = x) = \frac{e^{X\cdot\hat{\beta}}}{1 + e^{X\cdot\hat{\beta}}}$$

LDA:

The log-posterior of class k given an input x is:

$$C + \log \pi_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + x^T \mathbf{\Sigma}^{-1}\mu_k$$

where C is a constant which does not depend on k.

QDA:

The log-posterior of class k given an input x in QDA is:

$$C + \log \pi_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}_k^{-1}\mu_k + x^T \mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}x^T \mathbf{\Sigma}_k^{-1}x - \frac{1}{2}\log|\mathbf{\Sigma}_k|$$

where C is a constant which does not depend on k.