Introduction to Machine Learning – Generative Models

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

Reminder

- Next Tuesday (02/20) is a Monday Schedule

Plan for Today

- Quadratic Discriminant Analysis
- Naive Bayes

Warm Up: Classification Errors

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

Predicted

Actual		**0**	**1**
	0	30	12
	1	8	56

Calculate specificity, sensitivity, and precision for the model that produced this confusion matrix.

Generative Models

- Logistic Regression directly models $\Pr(Y = k|X = x)$
  - i.e., we model the conditional distribution of Y given the predictor(s) X
- Alternatively, we can model the distribution of predictors, X, separately for each response class. Then use Bayes Theorem to flip them into estimates for $\Pr(Y = k|X = x)$

LDA

LDA makes 2 assumptions:

- Observations within class are normally distributed
- All classes have common variance

LDA assigns an observation, $X = x$ to the class for which the discriminant is largest.

LDA discriminant function:

$$\delta_k(x) = \frac{x^T \hat{\mu}_k}{\Sigma} - \frac{\hat{\mu}_k^T \hat{\mu}_k}{2\Sigma} + \log(\hat{\pi}_k)$$

LDA

LDA makes 2 assumptions:

- Observations within class are normally distributed
- All classes have common variance

What if we relax this assumption?

LDA assigns an observation, $X = x$ to the class for which the discriminant is largest.

LDA discriminant function:

$$\delta_k(x) = \frac{x^T \hat{\mu}_k}{\Sigma} - \frac{\hat{\mu}_k^T \hat{\mu}_k}{2\Sigma} + \log(\hat{\pi}_k)$$

LDA

- Relax the LDA assumption that classes have uniform variance

- That means we now have

$$\Sigma \to \Sigma_k$$

LDA

- Relax the LDA assumption that classes have uniform variance

- That means we now have

$$\Sigma \rightarrow \Sigma_k$$

- If we plus this into Bayes we get:

$$\delta_k(x) = -\frac{x^T x}{2\Sigma_k} + \frac{x^T \mu_k}{\Sigma_k} - \frac{\mu_k^T \mu_k}{2\Sigma_k} - \frac{\log|\Sigma_k|}{2} + log\pi_k$$

QDA

LDA → ***Quadratic Discriminant Analysis***

- Relax the LDA assumption that classes have uniform variance
- That means we now have

$$\Sigma \rightarrow \Sigma_k$$

- If we plus this into Bayes we get:

$$\delta_k(x) = -\boxed{\frac{x^T x}{2\Sigma_k}} + \frac{x^T \mu_k}{\Sigma_k} - \frac{\mu_k^T \mu_k}{2\Sigma_k} - \frac{\log|\Sigma_k|}{2} + log\pi_k$$

Multiplying two x terms together → quadratic

LDA vs QDA

Bias-Variance Tradeoff

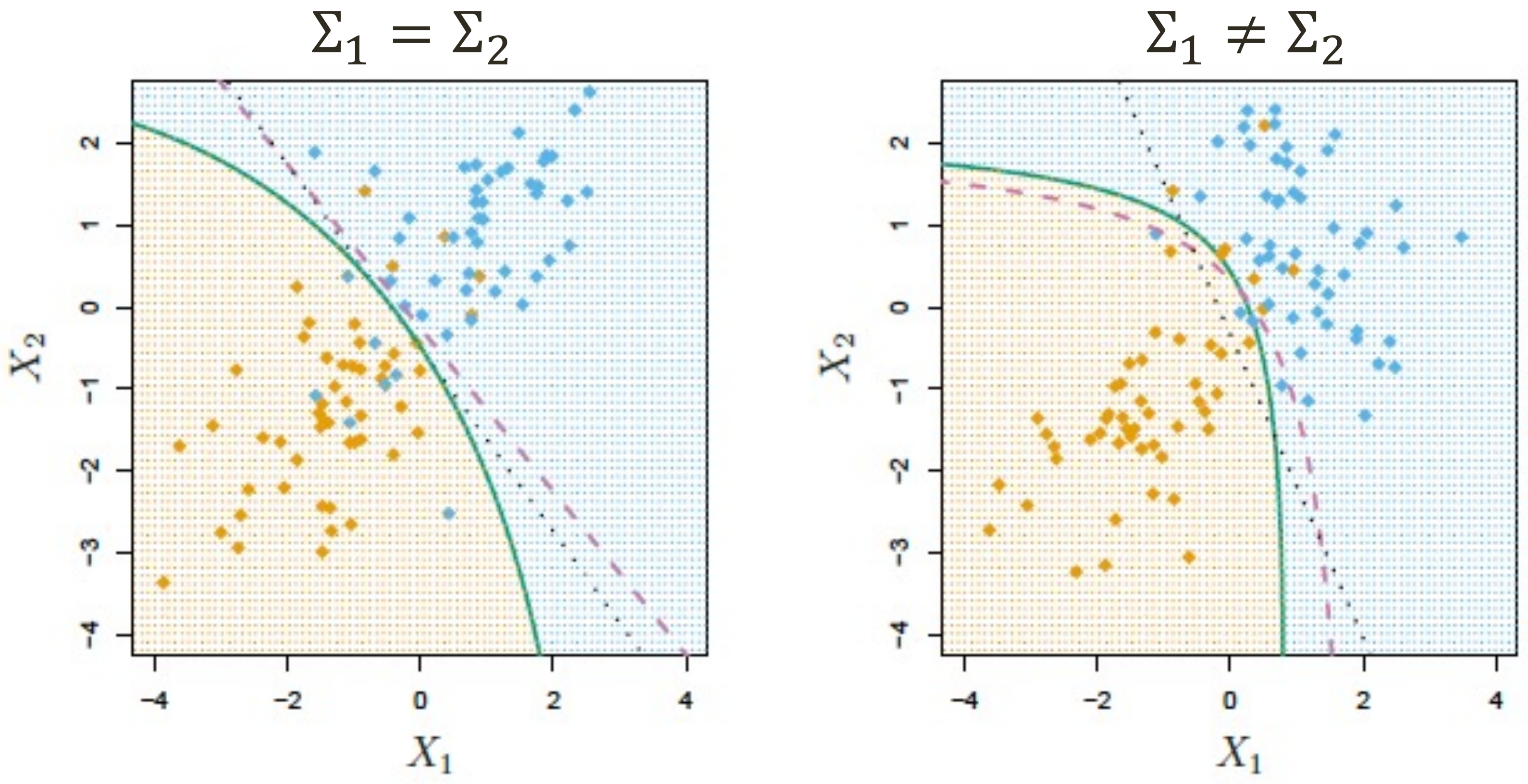
If we have p predictors….

- Estimating a covariance matrix requires estimating $\frac{p(p+1)}{2}$ parameters
- LDA assumes one covariance matrix, and is linear so there are Kp linear coefficients to estimate
- QDA estimates a covariance matrix for each class, K, for a total of $\frac{Kp(p+1)}{2}$ parameters to estimate

Which model has higher bias? Which has higher variance?

LDA vs QDA

Bias-Variance Tradeoff



- Purple dashed = Bayes
- Black dotted = LDA
- Green = QDA

Bayes Theorem

Easy to estimate! How?

Hard to estimate, many options

Remember…

Bayes Theorem

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- π_k is the ***prior probability*** that a randomly chosen observation come from the kth class
- $f_k(X) \equiv \Pr(X|Y = k)$ is the ***density function*** of X for an observation that comes from the kth class
- $\Pr(Y = k|X = x)$ is the ***posterior probability***, i.e. the probability that an observation belongs to the kth class given the predictor value *(x)* for the observation

Naive Bayes

Hard to estimate, many options

Remember…

Bayes Theorem

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- $f_k(X) \equiv \Pr(X|Y = k)$ is the ***density function*** of X for an observation that comes from the kth class
- LDA and QDA assume $f_k(x)$ is multivariate normal
- naive Bayes classifier assumes:
  - Within the kth class, the p predictors are independent

Naive Bayes

Assumes

Within the kth class, the p predictors are independent

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$$

f_{kj} is the density function for the jth predictor among observations in the kth class

Naive Bayes

Assumes

Within the kth class, the p predictors are independent

$$f_k(x) = f_{k1}(x_1)\times f_{k2}(x_2)\times\cdots\times f_{kp}(x_p)$$

f_{kj} is the density function for the jth predictor among observations in the kth class

- This assumption eliminates the need to estimate covariance (there is no covariance if everything is independent!)

In practice, do you expect all predictors to be independent?

Naive Bayes

Posterior probability:

$$\Pr(Y = k)\,|X = x) = \frac{\pi_k \times f_{k1}(x_1) \times f_{ks}(x_2) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times f_{ls}(x_2) \times \cdots \times f_{lp}(x_p)}$$

Options for estimating f_{kj}:

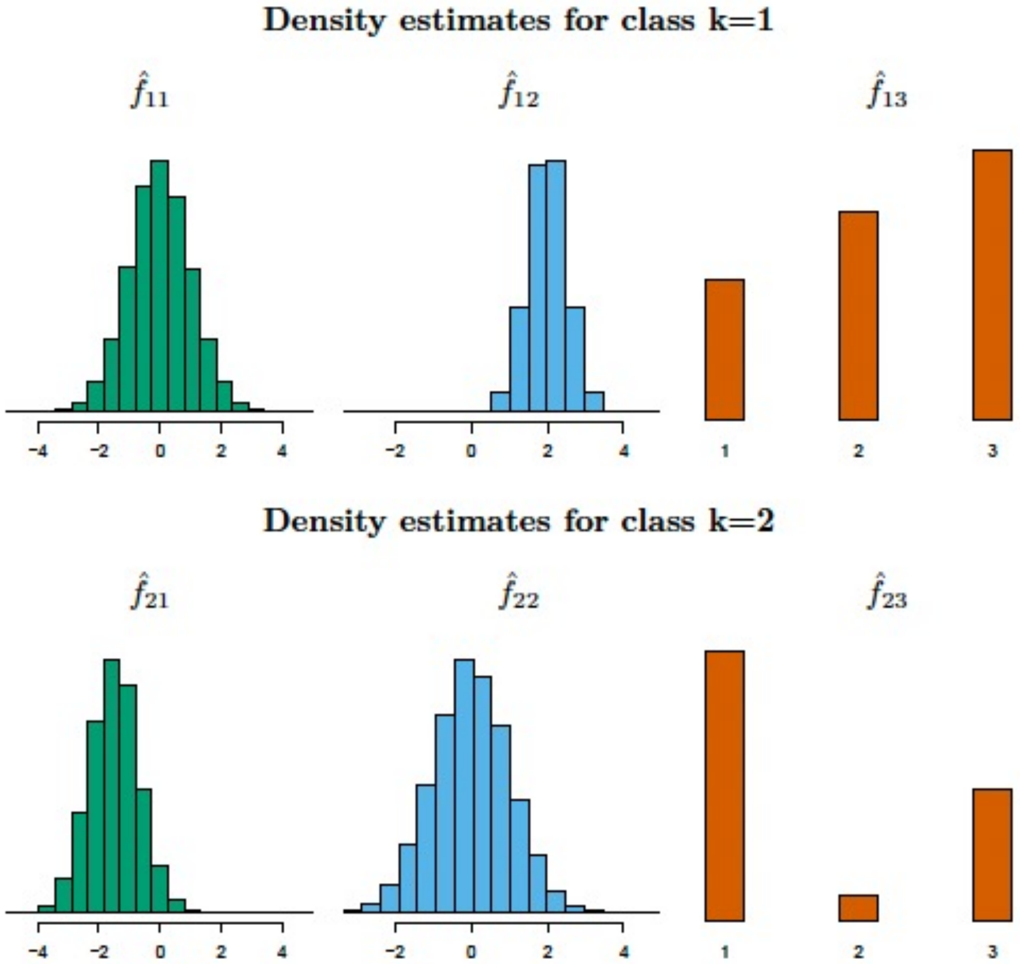
- If X_j is quantitative, assume univariate normal distributions for each predictor within each class
- If X_j is quantitative, use non-parametric estimate. Make a histogram for observations of the jth predictor within each class and estimate $f_{kj}(x_j)$ as the fraction of training observations in the kth class in the same histogram bin as x_j
- If X_j is qualitative, count the proportion of training observation for the jth predictor corresponding to each class

Naive Bayes

Example:

$$\Pr(Y = k) \mid X = x) = \frac{\pi_k \times f_{k1}(x_1) \times f_{ks}(x_2) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times f_{ls}(x_2) \times \cdots \times f_{lp}(x_p)}$$

- $p = 3, K = 2$
- The first two predictors are quantitative, third is qualitative
- $\hat{\pi}_1 = \hat{\pi}_2 = 0.5$



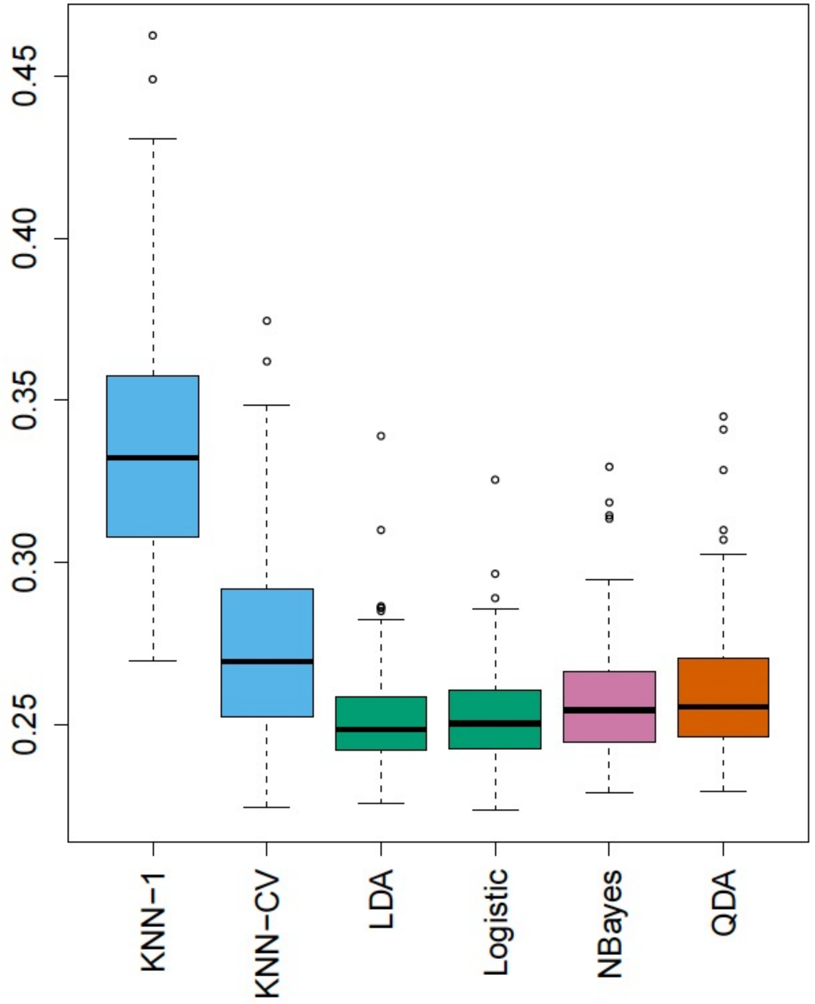
Predict the class of $x = \begin{bmatrix} 0.4 \\ 1.5 \\ 1 \end{bmatrix}$

$\hat{f}_{11}(0.4) = 0.368, \hat{f}_{12}(1.5) = 0.484,$
$\hat{f}_{13}(1) = 0.226, \hat{f}_{21}(0.4) = 0.030,$
$\hat{f}_{22}(1.5) = 0.130, \hat{f}_{23}(1) = 0.616$

Comparing Models

Scenario 1:

- K = 2, p = 2 (both quantitative), true relationship is linear
- 20 training observations in each class
- Observations are uncorrelated random normal variables

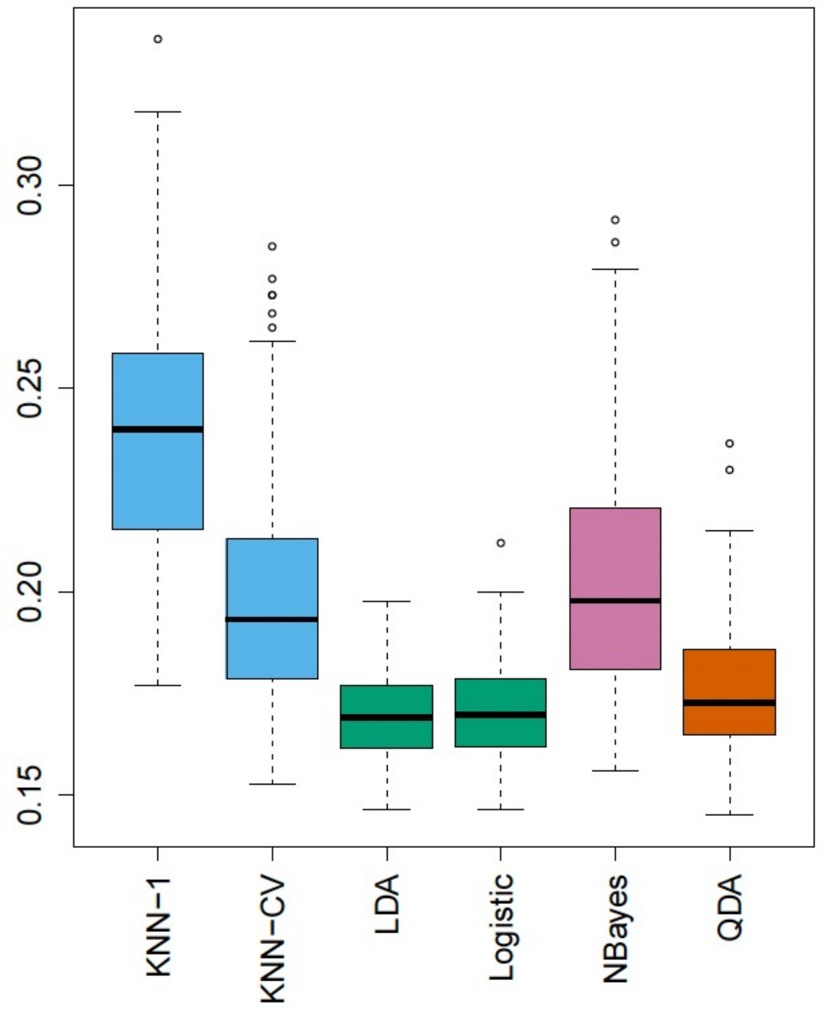


How do model performances compare?
Why do we see the ranking we see?

Comparing Models

Scenario 2:

- K = 2, p = 2 (both quantitative), true relationship is linear
- 20 training observations in each class
- Within each class, predictors have a correlation of -0.5

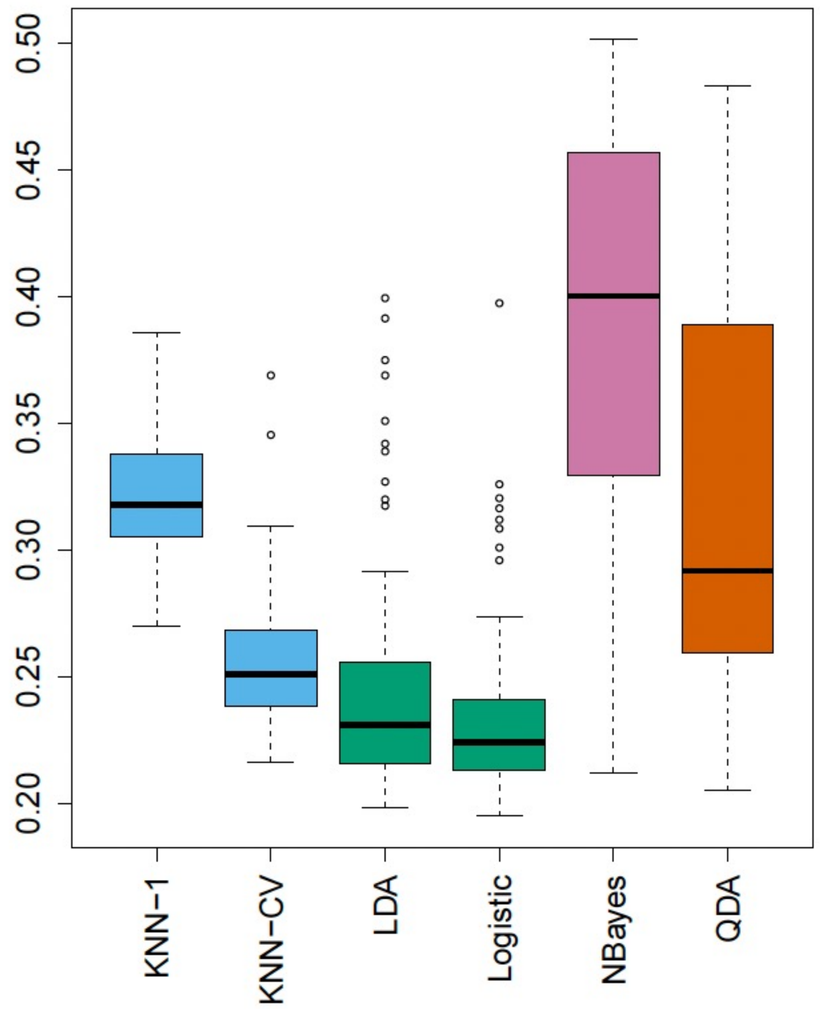


How do model performances compare?
Why do we see the ranking we see?

Comparing Models

Scenario 3:

- K = 2, p = 2 (both quantitative), true relationship is linear
- 50 training observations in each class
- Within each class, predictors have a correlation of -0.5
- X_1 and X_2 are generated from the t-distribution (similar to normal, but with longer tails)

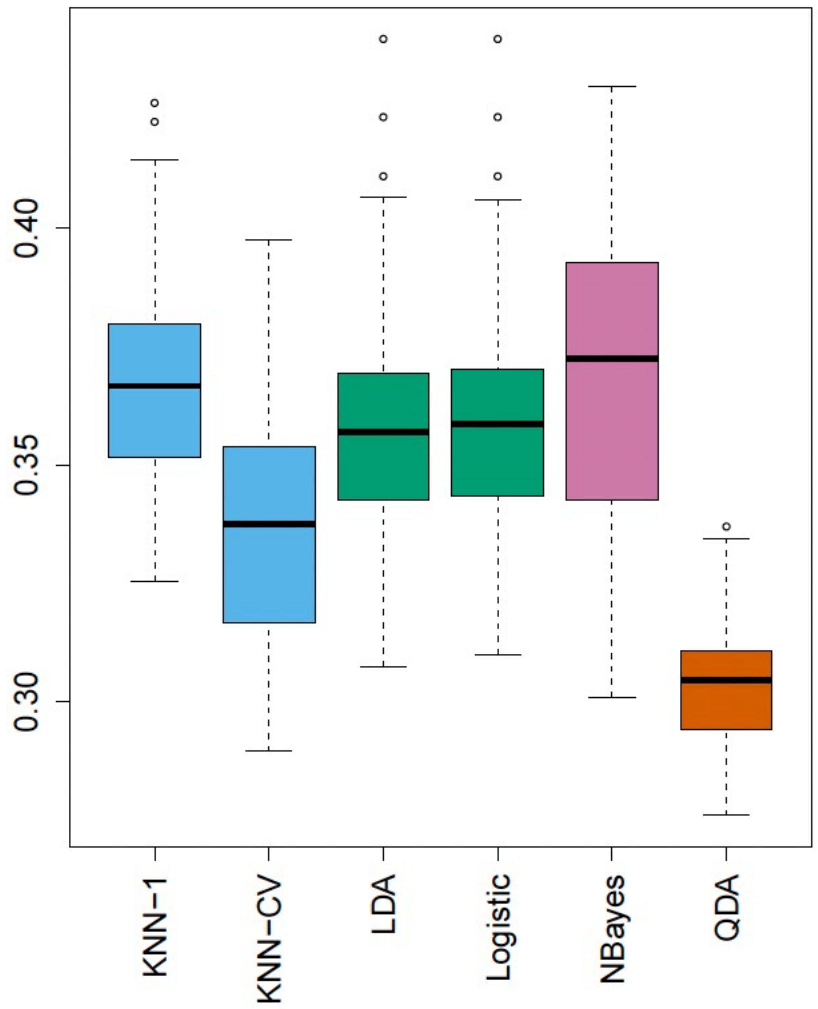


How do model performances compare?
Why do we see the ranking we see?

Comparing Models

Scenario 4:

- K = 2, p = 2 (both quantitative), true relationship is non-linear
- Within class 1, predictors have a correlation of 0.5
- Within class 2, predictors have a correlation of -0.5
- X_1 and X_2 are generated from the normal distribution

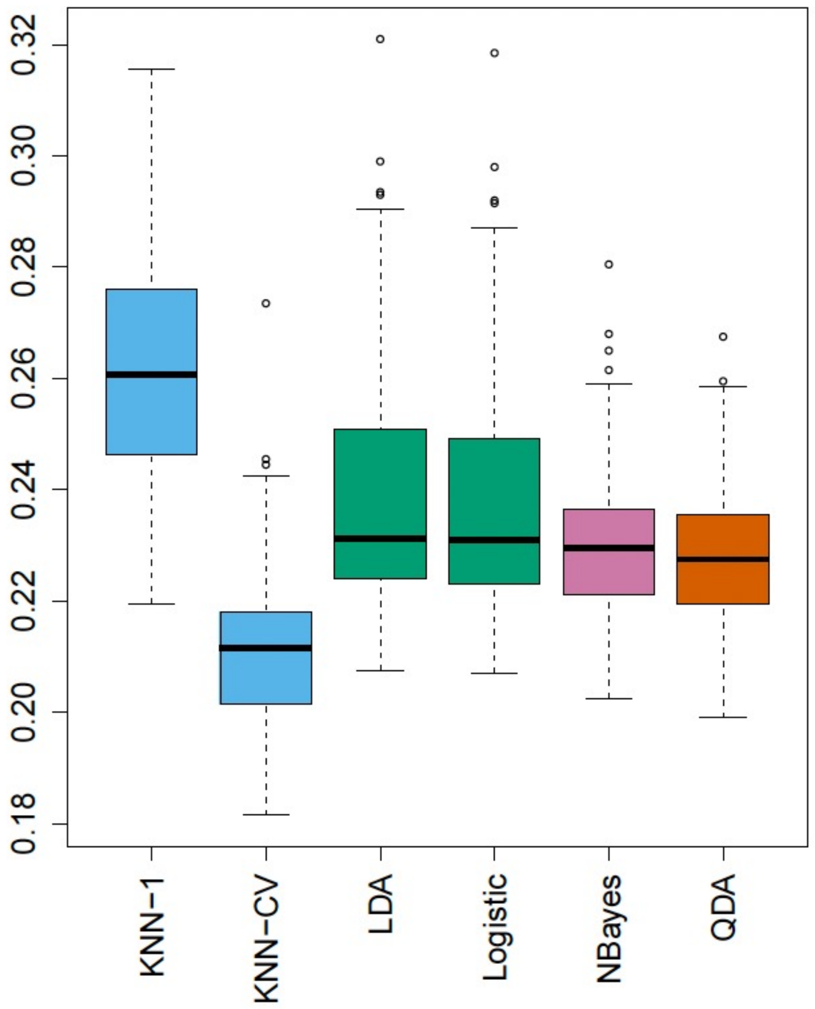


How do model performances compare?
Why do we see the ranking we see?

Comparing Models

Scenario 5:

- K = 2, p = 2 (both quantitative), true relationship is non-linear
- Responses were first generated from the normal distribution with uncorrelated predictors. Then responses were sampled from the logistic function applied to a complicated non-linear function of predictors

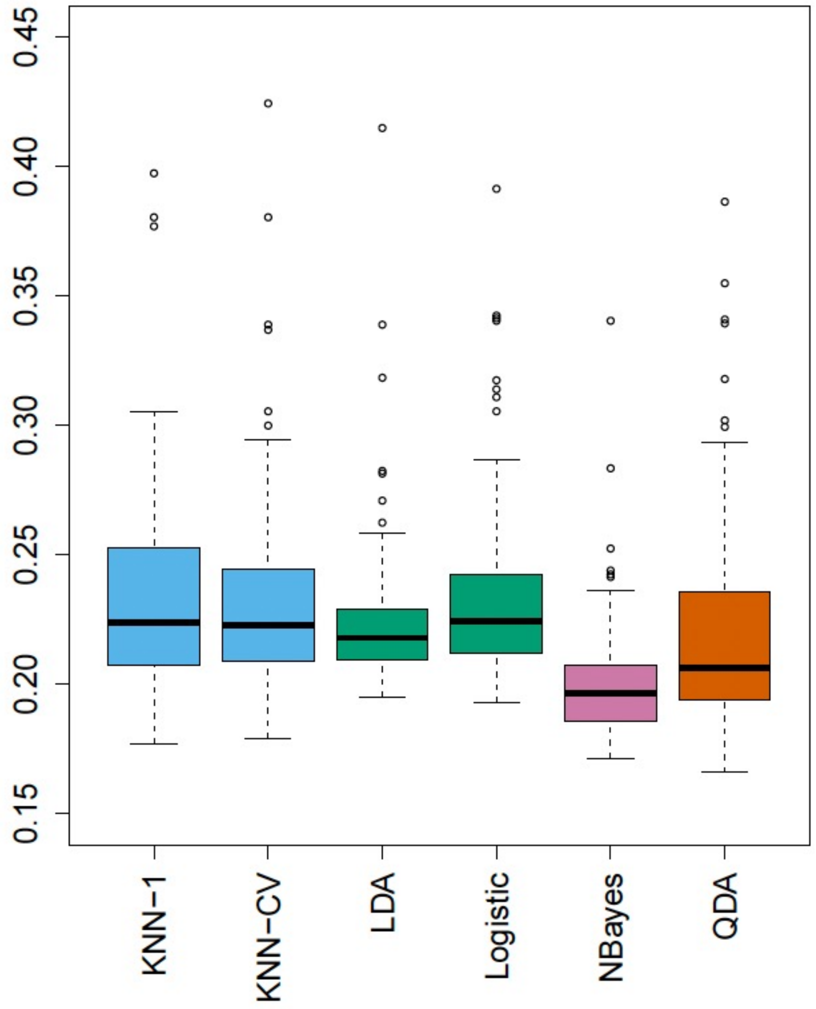


How do model performances compare?
Why do we see the ranking we see?

Comparing Models

Scenario 6:

- K = 2, p = 2 (both quantitative), true relationship is non-linear
- Responses were generated from the normal distribution with different covariance for each class
- 6 training observations in each class



How do model performances compare?
Why do we see the ranking we see?