Classification II

Lecture 5

Termeh Shafie

The Bayes in Naive Bayes

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

$$P(\text{category} \mid x_1, x_2, x_3) = \frac{P(x_1, x_2, x_3 \mid \text{category})P(\text{category})}{P(x_1, x_2, x_3)}$$

The Bayes in Naive Bayes

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

for computational efficiency

$$P(\text{diabetes} \mid \text{HS, O, HRH}) = \frac{P(\text{HS, O, HRH} \mid \text{diabetes})\ P(\text{diabetes})}{\cancel{P(\text{HS, O, HRH})}}$$

$$P(\text{diabetes}^c \mid \text{HS, O, HRH}) = \frac{P(\text{HS, O, HRH} \mid \text{diabetes}^c)P(\text{diabetes}^c)}{\cancel{P(\text{HS, O, HRH})}}$$

diabetes = high/low risk
HS = high sugar intake
O = Obese
HRH = high resting heart rate

The Bayes in Naive Bayes

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

$$P(\text{diabetes} \mid \text{HS, O, HRH}) \propto P(\text{HS, O, HRH} \mid \text{diabetes})\ P(\text{diabetes})$$

$$P(\text{diabetes}^c \mid \text{HS, O, HRH}) \propto P(\text{HS, O, HRH} \mid \text{diabetes}^c)P(\text{diabetes}^c)$$

diabetes = high/low risk
HS = high sugar intake
O = Obese
HRH = high resting heart rate

The Naive in Naive Bayes

features are **conditionally independent** given the class label

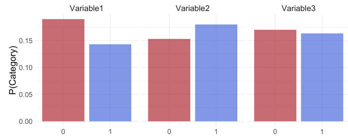
$$P(\text{HS}, \text{O}, \text{HRH}) = P(\text{HS}) \cdot P(\text{O}) \cdot P(\text{HRH})$$

diabetes = high/low risk
HS = high sugar intake
O = Obese
HRH = high resting heart rate

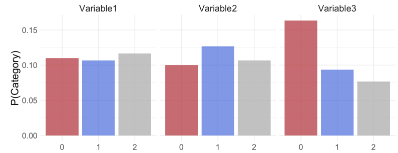
5

Naive Bayes

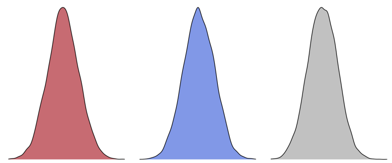
Bernoulli Naive Bayes



Categorical Naive Bayes

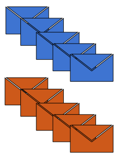


Gaussian Naive Bayes



6

Bernoulli Naive Bayes



spam	dear	lunch	viagra	money
0	0.25	0.46	0.01	0.14
1	0.32	0.05	0.53	0.67

$$P(\text{category} \mid x_1, x_2, \ldots, x_p) \propto \prod_{i=1}^{p} P(x_i \mid \text{category}) \cdot P(\text{category})$$

 [1,1,0,0]

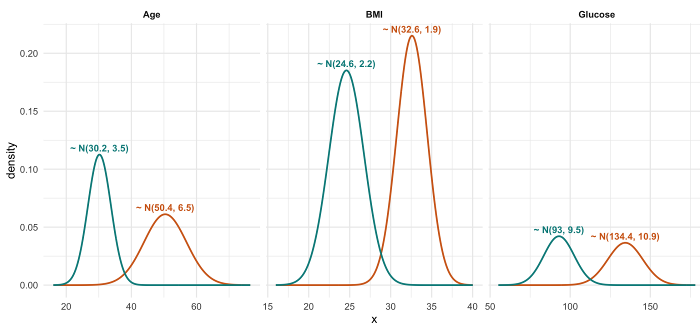
7

Gaussian Naive Bayes

predicting risk of diabetes

estimated Normal(μ, σ) distributions for low-risk and high-risk groups

prediction is based on which class makes the observed values more likely

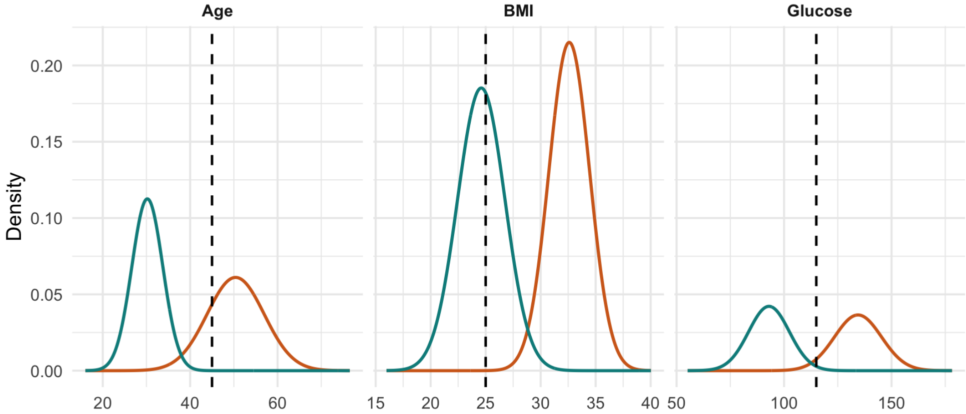


ID	BMI	Glucose	Age	Risk
1	22	82	25	low
2	24	90	30	low
3	23	88	28	low
4	26	95	35	low
5	31	125	45	high
6	33	135	50	high
7	35	142	55	high
8	30	120	42	high
9	28	110	33	low
10	34	150	60	high
⋮	⋮	⋮	⋮	⋮

8

Gaussian Naive Bayes

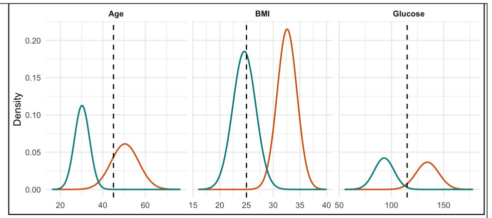
new observation:
Age = 45, BMI = 25, Glucose = 115



Gaussian Naive Bayes

class-conditional normal distribution:

$$p(x_j \mid y = c) = \frac{1}{\sqrt{2\pi\sigma_{c,j}^2}} \exp\left(-\frac{(x_j - \mu_{c,j})^2}{2\sigma_{c,j}^2}\right)$$



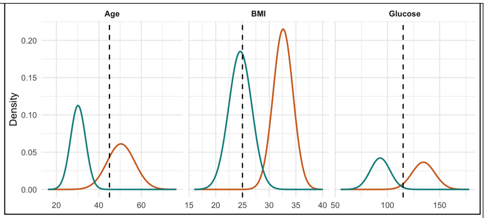
$$P(L) \times P(\text{Age} = 45 \mid L) \times P(\text{BMI} = 25 \mid L) \times P(\text{Glucose} = 115 \mid L)$$

$$P(H) \times P(\text{Age} = 45 \mid H) \times P(\text{BMI} = 25 \mid H) \times P(\text{Glucose} = 115 \mid H)$$

Gaussian Naive Bayes

class-conditional normal distribution:

$$p(x_j \mid y = c) = \frac{1}{\sqrt{2\pi\sigma_{c,j}^2}} \exp\left(-\frac{(x_j - \mu_{c,j})^2}{2\sigma_{c,j}^2}\right)$$



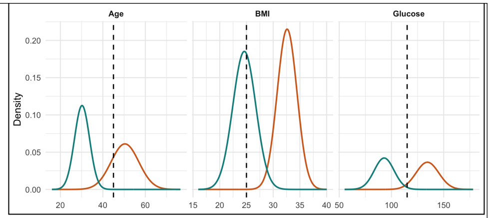
Feature	Likelihood (low)	Log-likelihood (low)
Age	0.00009406119	-9.2716
BMI	0.1633826	-1.8117
Glucose	0.004344250	-5.4389

Feature	Likelihood (high)	Log-likelihood (high)
Age	0.0415668350	-3.1805
BMI	0.0002329876	-8.3645
Glucose	0.0092573944	-4.6823

Gaussian Naive Bayes

class-conditional normal distribution:

$$p(x_j \mid y = c) = \frac{1}{\sqrt{2\pi\sigma_{c,j}^2}} \exp\left(-\frac{(x_j - \mu_{c,j})^2}{2\sigma_{c,j}^2}\right)$$



0.5

$$P(L) \times P(\text{Age} = 25 \mid L) \times P(\text{BMI} = 25 \mid L) \times P(\text{Glucose} = 115 \mid L)$$

$$P(H) \times P(\text{Age} = 25 \mid H) \times P(\text{BMI} = 25 \mid H) \times P(\text{Glucose} = 115 \mid H)$$

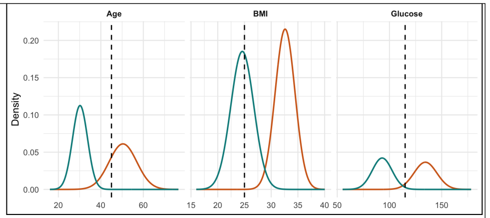
0.5

note: with log likelihood you sum up the probabilities and add the log of the prior

Gaussian Naive Bayes

class-conditional normal distribution:

$$p(x_j \mid y = c) = \frac{1}{\sqrt{2\pi\sigma^2_{c,j}}} \exp\left(-\frac{(x_j - \mu_{c,j})^2}{2\sigma^2_{c,j}}\right)$$



$$P(L \,|\, P(\text{Age} = 25,\ P(\text{BMI} = 25),\ P(\text{Glucose} = 115) \approx 0.43^*$$

$$P(H \,|\, P(\text{Age} = 25,\ P(\text{BMI} = 25),\ P(\text{Glucose} = 115) \approx 0.57^*$$

* true probabilities after normalization

Bayes Classifier vs. Naive Bayes

Warning: The Bayes classifier should not be confused with a Naive Bayes classifier!

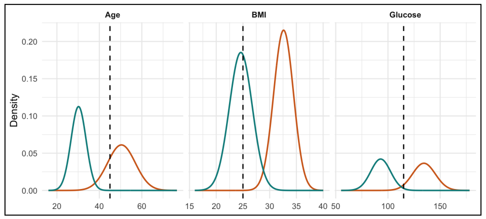
- **Bayes optimal classifier** (or **Bayes classifier**) is a theoretical construct, not a practical classification method
- It is defined as the classifier that has the smallest test error rate and **assumes we know P(category | predictors)**
- It's what we would ideally use if we knew the true data generating process

A probabilistic model-based approach to using Bayes classifier is:

1. Estimate the true distribution of test set from the training set
2. Use the Bayes optimal classifier for the estimated distribution

Bayes Classifier vs. Naive Bayes

Warning: The Bayes classifier should not be confused with a Naive Bayes classifier!



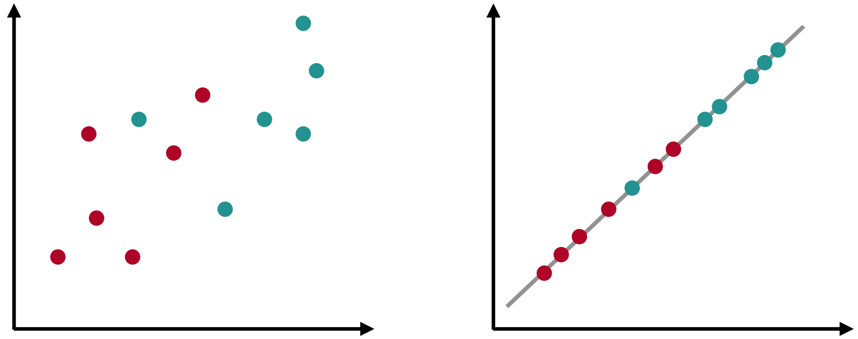
The Bayes optimal classifier uses the true joint distribution of Age, BMI, and Glucose to predict diabetes risk

Naive Bayes simplifies this by assuming the predictors are independent and modeling each with its own Normal(μ,σ) distribution.

Linear Discriminant Analysis (LDA)

- Model the distribution of predictors in each category separately
- Use **Bayes theorem** to flip things around and obtain P(category | predictors)
- Naive Bayes: features are ~~**conditionally independent given the class label**~~
- Now: **model the joint distribution of features given the class label**
  - assume distribution of the features within each category is normally distributed
  - assume covariances of the MVN distributions are equal for both classes
  - use the Bayes optimal classifier

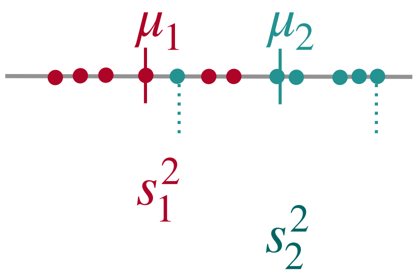
Linear Discriminant Analysis (LDA)



Linear Discriminant Analysis (LDA)

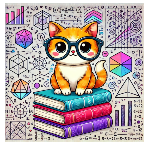
1) maximize the distance between the means

2) minimize the variation (s^2) within each category



$$\frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2}$$

LDA with One Predictor



- $f_k(x)$ is normal with following density in one dimension:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

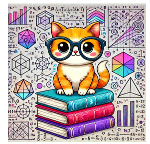
where μ_k and σ_k^2 are mean and variance of kth class and assume variances are equal

- Plug this into Bayes theorem

$$P_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

- The Bayes classifier assigns an observation to where the above is the largest which is equivalent to the largest discriminant score: $\delta_k(x) = x\frac{\mu_k}{\sigma^2} + \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$
- This is the linear discriminant classifier

LDA with Multiple Predictors



- Each class k has a multivariate normal distribution:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)$$

where p = number of predictors

$\boldsymbol{\mu}_k$= mean vector of class k

Σ= common covariance matrix (***same for all classes***)

- Plug this into Bayes theorem

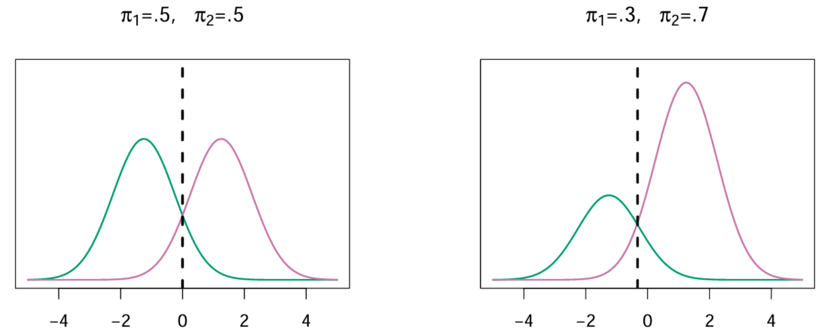
$$P_k(\mathbf{x})\frac{\pi_k f_k(\mathbf{x})}{\sum_{l=1}^{K} \pi_l f_l(\mathbf{x})}$$

- The Bayes classifier assigns an observation to the class with the largest discriminant score

$$\delta_k(\mathbf{x})\mathbf{x}^\top\Sigma^{-1}\boldsymbol{\mu}_k \bullet \frac{1}{2}\boldsymbol{\mu}_k^\top\Sigma^{-1}\boldsymbol{\mu}_k \bullet \log(\pi_k)\,.$$

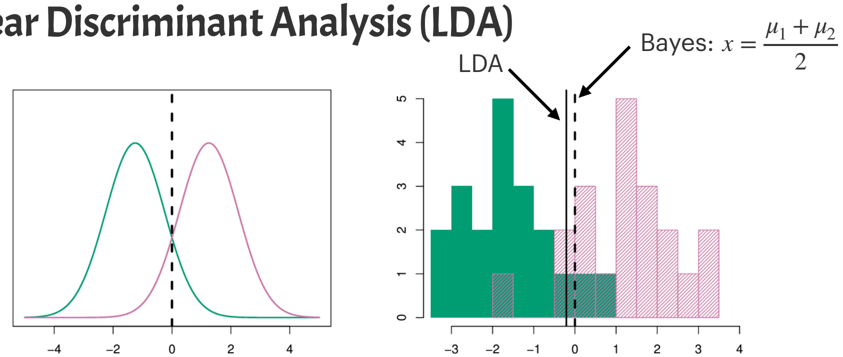
- This is the linear discriminant classifier (linear in $\mathbf{x}$).

Linear Discriminant Analysis (LDA)



- the dashed lined represents the Bayes decision boundary (Bayes Classifier)
- we classify a new point to which density is highest
- when priors are different, take them into account and compare $\pi_k f_k(x)$
- on the right, we favor the pink class the decision boundary has shifted to the left

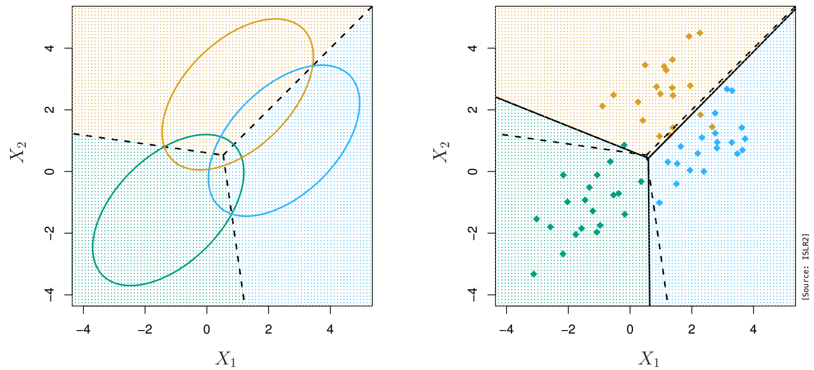
Linear Discriminant Analysis (LDA)



$\mu_1 = -1.5, \mu_2 = 1.5,$
$\sigma_1^2 = \sigma_2^2 = 1, \ \pi_1 = \pi_2 = 0.5$

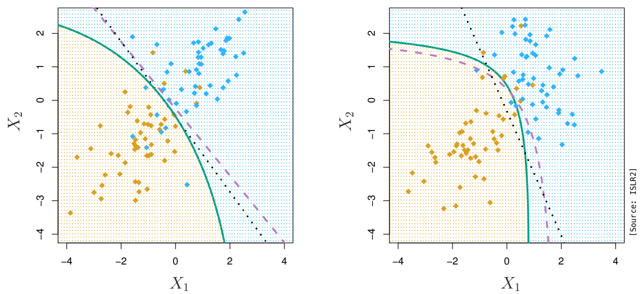
- typically we don't know these parameters
- in that case, we estimate them and plug them into the rule

LDA with Three Classes



Quadratic Discriminant Analysis (QDA)

does not assume a common covariance across classes for these MVNs



- LDA uses a linear decision rule (less flexible)
  - constrains such that it uses same covariance matrix for each class
- QDA uses a quadratic decision rule (more flexible)
  - allows each class k to have a different covariance matrix

KNN vs. Logistic Regression vs. LDA vs. QDA

- **KNN:** good when complex boundaries and n is sufficiently large
- **Logistic regression** and **LDA**: good when linear boundaries or p is big relative to n
  - LDA extends better to multi-class problems
  - LDA is more stable during estimation
  - Logistic regression is more robust to outliers
- **QDA**: good when quadratic (or moderately complex) boundaries and n is moderately big

Measuring Classification Performance

- Measure of classification performance is

  error rate = fraction of points that are classified incorrectly

- The **training error rate** is

$$\text{training error} = \frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \neq y_i)$$

- The (expected) **test error rate** is given by $\mathbf{E}\left(I(\hat{Y}_0) \neq Y_0\right)$
- We have to construct $\hat{f}$ to **minimize the test error rate**

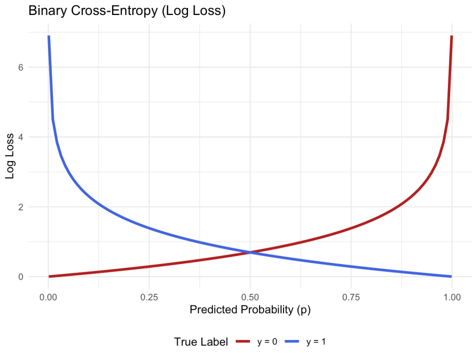
  $\Longrightarrow$ we need a **loss function** $L(\hat{y}, y)$ for penalizing errors in $\hat{y} = \hat{f}(x)$ when truth is y

- Strongly contingent on application

Binary Cross Entropy

$$l(\beta_0, \beta_1) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i))$$

`Loss function`



Assessing Model Performance

- Did it make the correct prediction?
  - accuracy
  - sensitivity
  - specificity
- How well does to perform in distinguishing classes correctly?

Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix: Accuracy

correct predictions

True Positive (TP) + True Negative (TN)

all predictions

False Negative (TN) + False Positive (FP) +

True Positive (TP) + True Negative (TN)

How often is the model correct?

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix: Sensitivity/Recall

correctly predicted positives

True Positive (TP)

actual positives

False Negative (TN) + True Positive (TP)

How often is the model correct for Positive Cases?

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix: Specificity

correctly predicted negatives

True Negative (TN)

actual negatives

True Negative (TN) + False Positive (FP)

How often is the model correct for Negative Cases?

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix: Precision

$$\frac{\text{correctly predicted positives: True Positive (TP)}}{\text{all positives: False Positive (FP)} + \text{True Positive (TP)}}$$

How many of the predicted Positives are correct?

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix: F1 Score

$$\frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times precision \times recall}{precision + recall}$$

Combination of
Precision (how often predicted positives ARE positive) and
Recall (how often we correctly predict actual positives)

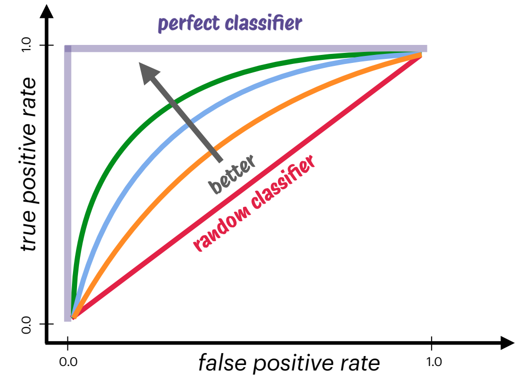
		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (TN)
	Negative	False Positive (FP)	True Negative (TN)

ROC AUC

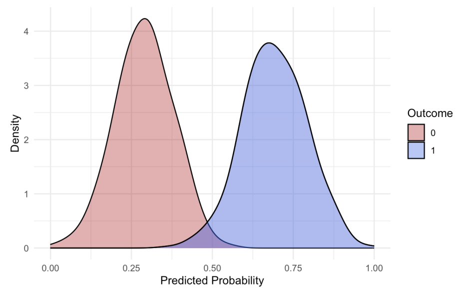
$$\frac{\text{correctly predicted positives: True Positive (TP)}}{\text{actual positives: False Negative (TN)} + \text{True Positive (TP)}}$$

$$1 - \frac{\text{correctly predicted negatives: True Negative (TN)}}{\text{actual negatives: True Negative (TN)} + \text{False Positive (FP)}}$$

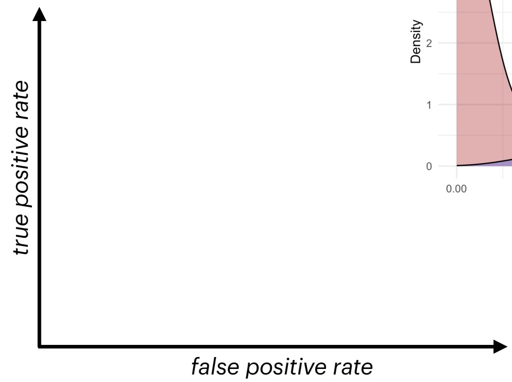
ROC AUC

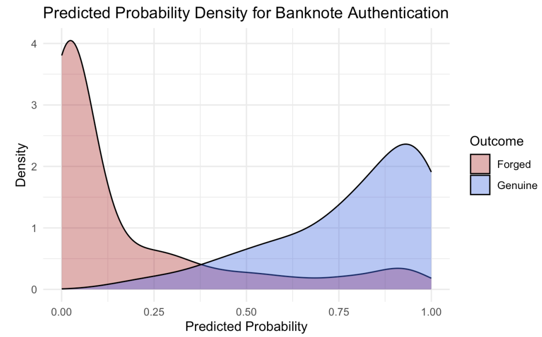


ROC AUC



ROC AUC





more realistic...

This Week's Practical

Classification and Model Evaluation

spam or ham text message