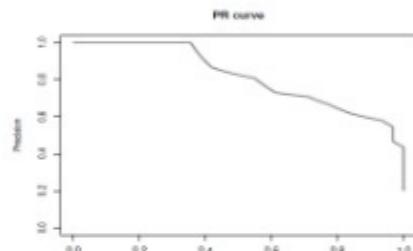


Precision-recall curves are often used for imbalanced applications

- Precision ($\frac{TP}{P}$) against the recall ($\frac{TP}{TP+FP}$).
- To plot a precision-recall curve we can use the package PRROC.

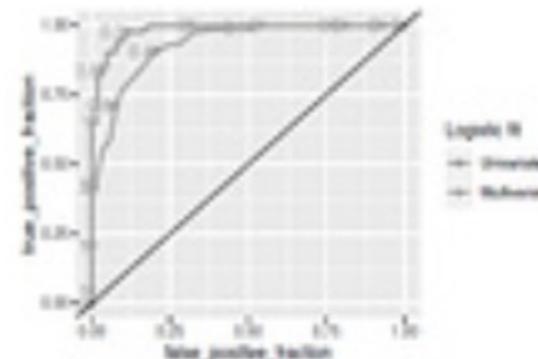
```
library(PRROC)

PRROC_obj <- pr.curve(scores.class0 = imbalanced_heights$mu_hat,
                      weights.class0=imbalanced_heights$y,
                      curve=TRUE)
plot(PRROC_obj, auc.main=FALSE , color=2)
```



Area under the ROC curve

- The area-under the ROC curve (AUC) captures the overall performance.
- AUC = 0.5 : random
- AUC = 1 : perfect



Problem 8 (2 credits)

One sample of only individuals is considered. It will be assumed we consider one variable consisting of 200000 adult human males, that a Gaussian normal distribution follows and represents a percentage of 50 percent (50% men, 50% women). In front of the hypothesis test, the researcher has a prior information of H_0 is true. In the other hand the prior information is H_0 is false. For such assumptions, the test is operated. Please consider that now H_0 is true and the value H_0 is false for rejecting H_0 is α . What is the power of the test if the rejection threshold is 0.05? How many false positives do we expect? If you implement an R code, consider some 1000000 data points to make consider.

$$\begin{aligned} H_0 & \text{ rejected} \\ \Downarrow & \\ p\text{-Value} & = \alpha \\ \Downarrow & \\ P_{H_0} (H_0 \text{ rejected}) & \leq \alpha \end{aligned}$$

- $E[\text{false positive}] = \frac{5000}{n} \cdot 0.05 = 250 \checkmark$
number of cases where H_0 is true
probability that H_0 is rejected although H_0 true
- $E[\text{false positive}] = 250 \checkmark$

Confusion Matrix

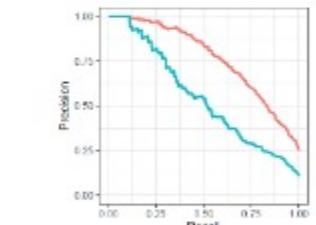
		predicted class	
		positive	negative
true class	positive	TP	FN
	negative	FP	TN

classifiers claims sample is positive (e.g. sick), and it's correct

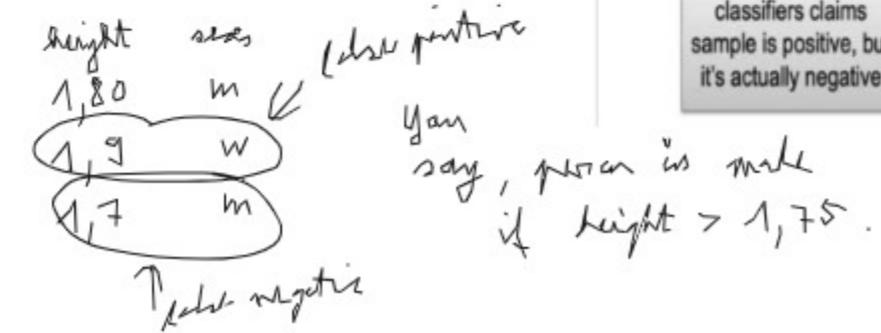
classifiers claims sample is negative, but it's actually positive

classifiers claims sample is positive, but it's actually negative

classifiers claims sample is negative, and it's correct



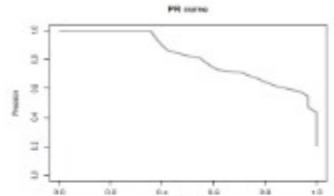
Based on the provided curves, is it possible to determine which of the two models has superior performance?



Precision-recall curves are often used for imbalanced applications

- Precision ($\frac{TP}{P}$) against the recall ($\frac{TP}{TP+FP}$).
- To plot a precision-recall curve we can use the package PRROC.

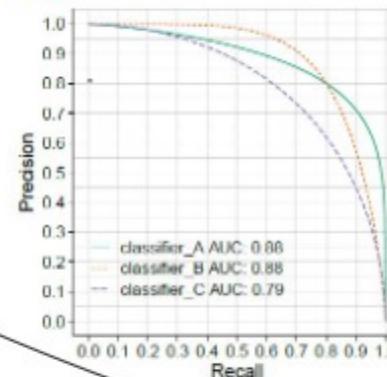
```
library(PRROC)
PRROC_obj <- pr.curve(scores.class0 = imbalanced_heights@mu_hat,
                      weights.class0=imbalanced_heights$y,
                      curve=TRUE)
plot(PRROC_obj, auc.main=FALSE , color=2)
```



positive = criminal

Problem 13 (1 credit)

A police department uses the help of a machine learning classifier to find criminals. They want to ensure that less than 10% of individuals classified as criminal are actually innocent. Which one of the three classifiers A, B or C shown in the Precision-Recall curves do you recommend to use? Justify your choice.



grün Kurve = best Match

da die grüne Kurve ab Precision von 0.8 den höchsten Recall hat
Hoher Recall -> niedriges FP -> wenige unschuldige werden als Kriminelle verurteilt

Classification performance metrics

- The sensitivity refers to the fraction of actual positives that is predicted to be positive:

$$\text{Sensitivity} = \frac{TP}{P} = \frac{TP}{TP+FN}$$

The sensitivity is also referred to as "recall", "true positive rate", or "power".

- The specificity refers to the fraction of actual negatives that is predicted to be negative:

$$\text{Specificity} = \frac{TN}{N} = \frac{TN}{TN+FP}$$

The specificity is also known as "true negative rate" or "sensitivity of the negative class"

- The precision refers to the fraction of predicted positives that are indeed positives:

$$=\text{Precision} = \frac{TP}{TP+FP}$$

Rezult

Problem 7

The right table shows a scatter plot with data points for 5 and 10.



Given that the two clusters defined by x and y in the figure are well-separated according to a reasonable clustering algorithm, which of the given points is closest to the center of the 5 cluster?

- A $(2, 2)$ B $(3, 3)$ C $(4, 4)$
- D $(5, 5)$ E $(6, 6)$
- F $(7, 7)$ G $(8, 8)$

Justify your answer.

You _____ should be found in _____.

abhang'g' yea!

Problem 7 (2 credits)

You have a dataset named `dt`, which includes information on cars from the `mtcars` dataset in R. The dataset contains information on miles per gallon (mpg) and transmission type (automatic or manual) of the cars.

mpg	transmission
21	Manual
21	Manual
23	Manual
21	Automatic
19	Automatic
18	Automatic
...	...

A colleague intends to investigate whether there is a significant difference in the median miles per gallon between cars with automatic and manual transmissions using a permutation test. The colleague visualizes a histogram of the test statistics across the permuted datasets alongside the value of the statistic T in the original unpermuted dataset (T reference). Examine the plot and the code used.

```
1 # Permutation function
2 perm_test <- function(data, n_permutations = 1000) {
3   # get reference statistics
4   T_ref <- data[transmission == "Manual", median(mpg)] -
5     data[transmission == "Automatic", median(mpg)]
6
7   # get permutation statistics
8   T_star <- sapply(1:n_permutations, function(x){
9     # shuffle groups
10    data[, transmission := sample(transmission), by='transmission']
11    # compute statistics
12    data[transmission == "Manual", median(mpg)] -
13      data[transmission == "Automatic", median(mpg)] })
14
15  g <- ggplot(data = data.table(T_star = T_star), aes(T_star)) +
16    geom_histogram() +
17    geom_vline(aes(xintercept=T_ref, color="T_ref")) + xlim(0, 10)
18  print(g)
19  return(list(T_ref=T_ref, T_star=T_star))
20 }
```

Problem 7 (2 credits)

You have a dataset named `dt`, which includes information on cars from the `mtcars` dataset in R. The dataset contains information on miles per gallon (mpg) and transmission type (automatic or manual) of the cars.

mpg	transmission
21	Manual
21	Manual
23	Manual
21	Automatic
19	Automatic
18	Automatic
...	...

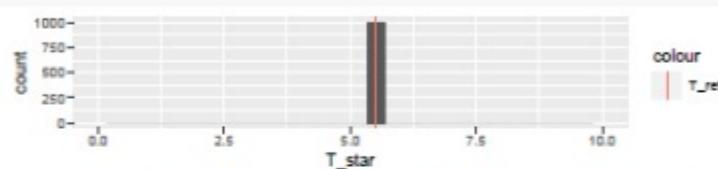
Jetzt für T_{obs}
 Durchschnittsverbrauch der
 Gangschaltautos
 • Durchschnittsverbrauch der
 Automatikschaltautos
 H_0 : Gangschaltautos verbrauchen weniger

A colleague intends to investigate whether there is a significant difference in the median miles per gallon between cars with automatic and manual transmissions using a permutation test. The colleague visualizes a histogram of the test statistics across the permuted datasets alongside the value of the statistic T in the original unpermuted dataset (T reference). Examine the plot and the code used.

```
# Permutation function
perm_test <- function(data, n_permutations = 1000) {
  # get reference statistics
  T_ref <- data[transmission == "Manual", median(mpg)] -
    data[transmission == "Automatic", median(mpg)]

  # get permutation statistics
  T_star <- sapply(1:n_permutations, function(x){
    # shuffle groups
    data[, transmission := sample(transmission, by="transmission")]
    # compute statistics
    data[transmission == "Manual", median(mpg)] -
      data[transmission == "Automatic", median(mpg)] })

  g <- ggplot(data = data.table(T_star = T_star), aes(T_star)) +
    geom_histogram() +
    geom_vline(aes(xintercept=T_ref, color="T_ref")) + xlim(0, 10)
  print(g)
  return(list(T_ref=T_ref, T_star=T_star))
}
```



- The code has a mistake. Clearly state the line of code that is wrong. Explain why it leads to the obtained histogram. Correct the line of code.
- Assume the code is now correct. Give the formula to compute the P-value estimate using T_{ref} and T_{star} in this particular case. No R code is required.

b) $T_{ref} = T_{obs}$
 $r = \#\{i \mid T_{star}[i] \geq T_{obs}\}$

$$\hat{P} = \frac{r+1}{m+1}$$

length(T_{star})
 //
 n - Permutation

Permutation testing: P-value

permutation

1/22 < >

- Let

- m be the number of random (Monte Carlo) permutations
- $r = \#\{T^* \geq T_{obs}\}$ be the number of these random permutations that produce a test statistic greater than or equal to that calculated for the actual data.

- The estimated one-sided P-value, \hat{P} is:

$$\hat{P} = \frac{r+1}{m+1}$$

- Permutation P-values should never be zero.

Falls $T_{obs} \geq T_{ref}$
 mit H_0
 abweichen
 Falls $T_{obs} < T_{ref}$, no r klein!
 \hat{P} klein
 \hat{P} klein
 H_0 wird abweichen

Permutation testing: P-value

permutation

1/22 < >

- Let

- m be the number of random (Monte Carlo) permutations.
- $r = \#\{T^* \geq T_{obs}\}$ be the number of these random permutations that produce a test statistic greater than or equal to that calculated for the actual data.

- The estimated one-sided P-value, \hat{P} is:

$$\hat{P} = \frac{r+1}{m+1}$$

- Permutation P-values should never be zero.

Problem 3 (10 points)

Die obigen Daten sind aus einer Studie über die Wirkung von zwei verschiedenen Arzneimitteln auf die Anzahl der Ausscheidungen von bestimmten Stoffen im Urin untersucht.

H_0 : Freund 1 und 2 Ausscheidungen unabhängig miteinander

T_{obs} = Anzahl der Beobachtungen, wo Freund 1 und 2 gleiche Tasten haben

$$T_{\text{obs}} = 4$$

Falls T_{obs} groß, dann wird H_0 abgelehnt

$$v = \#\{T^* \geq T_{\text{obs}}\}$$

W.L.G.

$$\hat{p} = \frac{v+1}{m+1} \quad \text{if } v > 0$$

H_0 wird abgelehnt



$\text{dt}[, \text{friend2} := \text{sample}(\text{friend2})]$

$t_{-\text{obs}}[i] = \text{dt}[\text{friend} == \text{friend2}, .N]$

$T^*[i] =$ Wert der Teststatistik
für die i-te Permutation

$r = 0;$

for i in $1:N_perm$ {
 if ($T_{-\text{obs}}[i] \geq T_{\text{ref}}$) {

$r = r + 1;$

}

$p = (r+1) / (N_perm + 1)$