Evaluation Metrics for Blocking/Entity Resolution

STA 325: Homework 2

General instructions for homeworks: Your code must be completely reproducible and must compile. No late homeworks will be accepted.

Reading Read the paper Binette and Steorts (2022) to get an overview of entity resolution. You'll want to refer to this during the course of the semester as it's meant to be a quick reference regarding the concepts that we will be covering. For more details, refer to the book by Christen (2012).

Advice: Start early on the homeworks and it is advised that you not wait until the day of as these homeworks are meant to be longer and treated as case studies.

Commenting code Code should be commented. See the Google style guide for questions regarding commenting or how to write code https://google.github.io/styleguide/Rguide.xml.

R Markdown Test

0. Open a new R Markdown file; set the output to HTML mode and "Knit". This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission.

Total points on assignment: 2 (reproducibility) + 23 points for the assignment = 25 total points.

1. (4 points) What are the four main challenges of entity resolution?

a. Costly manual labeling: Vast amounts of manually labeled data are required for supervised learning and evaluation.
b. Limited treatment of uncertainty: Given inherent uncertainties, it's important to output predictions with confidence regions.
c. Computational efficiency: Approximations are required to avoid quadratic scaling while ensuring there is minimal impact on accuracy.
d. Unreliable evaluation: Standard evaluation methods return imprecise estimates of performance.

2. (4 points, 1 point each) Suppose there are 10 records in a data set. a.) What are the total number of brute-force comparisons needed to make all-to-all record comparisons?

```
# Total number of all to all record comparisons
choose(10, 2)
```

```
## [1] 45
```

b.) Repeat this for 100 records, 1000 records, 10,000 records.

```r
choose(100, 2)
```

```
## [1] 4950
```

```r
choose(1000, 2)
```

```
## [1] 499500
```

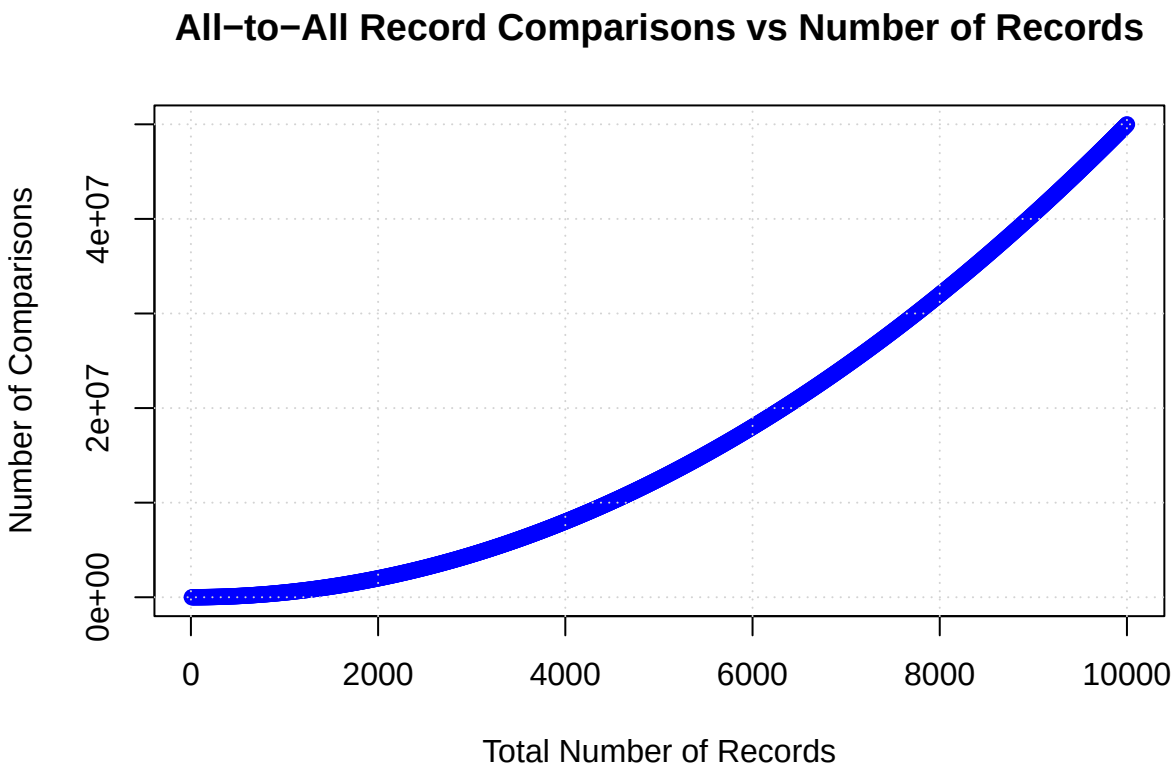
```r
choose(10000, 2)
```

```
## [1] 49995000
```

c.) As the number of records increases, comment on the growth rate of the number all-to-all record comparisons? Hint: plot all-to-all record comparisons versus total number of records What do you observe about the number of comparisons that need to be made?

The growth rate of the number of all-to-all record comparisons is quadratic. As the number of records increases, the number of comparisons grows quickly, making this approach computationally expensive for large data sets.

```r
# Define the number of records
n_records <- seq(10, 10000, by=10)

# Calculate the number of comparisons: n choose 2
comparisons <- (n_records * (n_records - 1)) / 2

# Plotting
plot(n_records, comparisons, type="b", col="blue", pch=19,
     xlab="Total Number of Records", ylab="Number of Comparisons",
     main="All-to-All Record Comparisons vs Number of Records")
grid()
```



3. (9 points) Consider the following record linkage data set with 1,000,000 total records that are matched between two databases. Assume that 500,000 are true matches. Assume a classified (or method) finds 600,000 record pairs as matches, and of these 400,000 correspond as true matches. The number of TP + FP + TN + FN = 50,000,000.

a. (4 points) Given the information above, find the following information in the confusion matrix: TP, FP, TN, and FN.

TP = 400,000 (given)

FP = 200,000 (predicted matches - TP)

FN = 100,000 (true matches - TP)

TN = 300,000 (non matches - FP)

b. (1 point) Calculate the accuracy. Comment on the reliability of this metric for this problem.

The accuracy formula can be displayed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{400000 + 300000}{1000000} = 0.7$$

Accuracy can be quite reliable for this problem since there are an equal number of matches and non-matches, leading to no class imbalance.

c. (1 point) Calculate the precision.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{400000}{600000} = 0.667$$

d. (1 point) Calculate the recall.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{400000}{500000} = 0.8$$

e. (1 point) Calculate the f-measure.

$$\text{F-measure} = 2 \times \frac{0.667 \times 0.8}{0.667 + 0.8} = 0.3637$$

f. (1 point) Comment on the reliability of the precision, recall, and f-measure for this problem.

For this problem, precision is quite reliable since there are few false negatives. Recall is also reliable but less than precision since information about the false positives is lost.F-measure is reliable since it provides a balanced trade-off between the two.

4. (6 points) We will revisit the Italian Survey on Household and Wealth (SHIW) from class, which is a sample survey 383 households conducted by the Bank of Italy every two years (2008 and 2010). The data set is anonymized to remove first and last name (and other sensitive information).

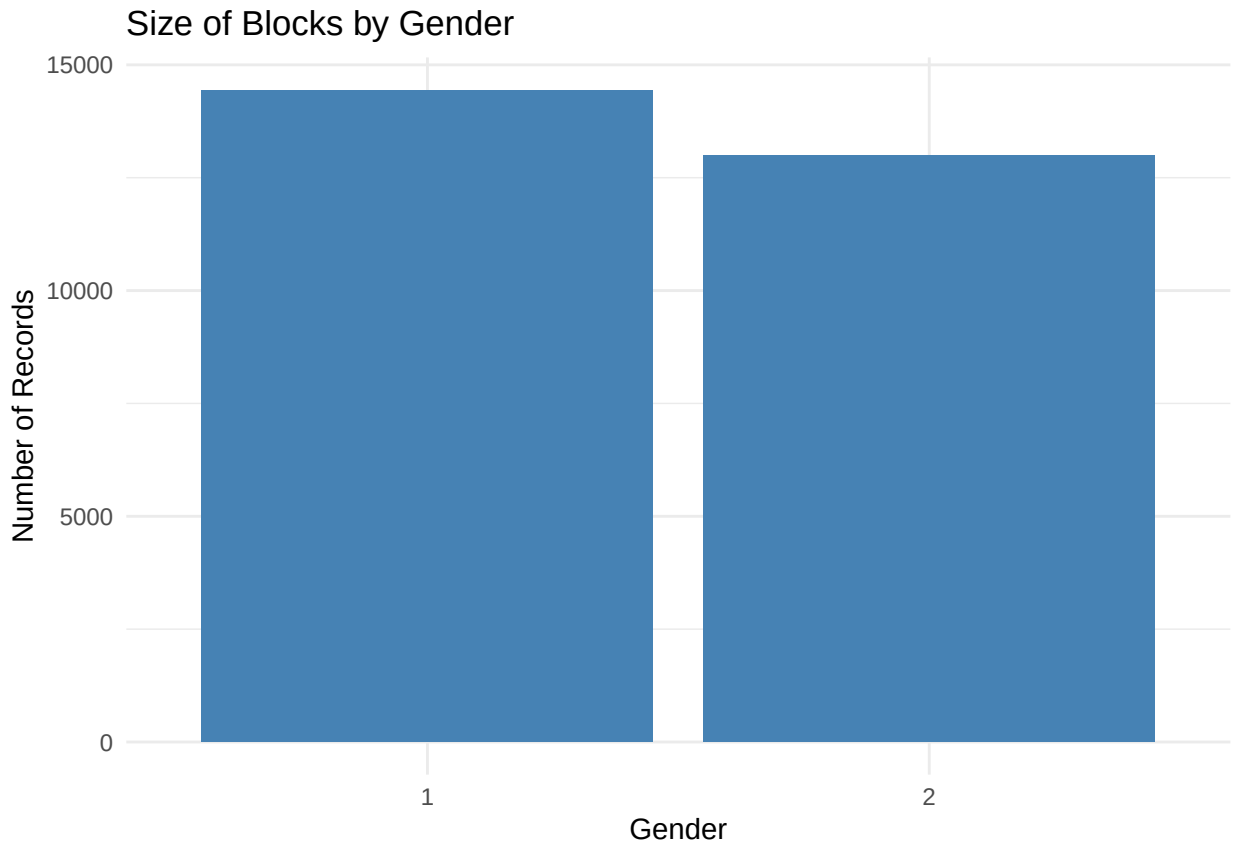
a. (0 points) Please load the data set in the way that we did in class and block based upon gender.

```
if (!require("italy")) {
  install.packages("italy")
}
if (!require("assert")) {
  install.packages("assert")
}
library(italy)
library(assert)
data(italy08)
data(italy10)

id08 <- italy08$id
id10 <- italy10$id
id <- c(italy08$id, italy10$id) # combine the id
italy08 <- italy08[-c(1)] # remove the id
italy10 <- italy10[-c(1)] # remove the id
italy <- rbind(italy08, italy10)

#block by gender
blockByGender <- italy$SEX
recordsPerBlock <- table(blockByGender)
```

b. (1 point) Plot the size of the blocks and comment on how many their are and their relative size.

There are 2 blocks of with gender 1 having a few more records than gender 2.

```r
if (!require("ggplot2")) {
  install.packages("ggplot2")
}
library(ggplot2)

ggplot(data = as.data.frame(recordsPerBlock), aes(x = names(recordsPerBlock), y = recordsPerBlock)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Gender", y = "Number of Records", title = "Size of Blocks by Gender") +
  theme_minimal()
```



c. (1 point) Calculate the reduction ratio and interpret its meaning.

```r
(choose(nrow(italy), 2) - sum(choose(recordsPerBlock, 2))) / choose(nrow(italy), 2)
```

```
## [1] 0.4986234
```

The reduction ratio indicates that the comparison space has been reduced by almost 50%.

d. (2 points) Calculate the precision and recall. Interpret the meaning of each.

```r
precision <- function(block.labels, IDs) {
  # Confusion matrix is a contingency table
  ct = xtabs(~block.labels + IDs)
```

```
  # Number of true positives
  TP = sum(choose(ct, 2))

  # Number of positives = TP + FP
  P = sum(choose(rowSums(ct), 2))

  return(TP / P)
}

# Define the recall function
recall <- function(block.labels, IDs) {
  # Confusion matrix is a contingency table
  ct = xtabs(~IDs + block.labels)

  # Number of true positives
  TP = sum(choose(ct, 2))

  # Number of true links = TP + FN
  TL = sum(choose(rowSums(ct), 2))

  return(TP / TL)
}

print(paste("Precision:", (precision(italy$SEX, id))))
```

```
## [1] "Precision: 3.59972673240546e-05"
```

```
print(paste("Recall:", (recall(italy$SEX, id))))
```

```
## [1] "Recall: 0.911310881524218"
```

The precision says that 0.003% of record pairs that are classified as matches correspond to true matches, while the rest correspond to false matches. The recall days that 91.1% of the true matching record pairs are correctly classified as matches.

e. (1 point) Would this be a reasonable approach for blocking. Explain.

This approach for blocking is relatively inefficient because it only reduces the number of comparisons by 50%. Furthermore, this blocking approach assumes that individuals of the same gender are likely to have similar attributes. This is untrue since attributes like employment, education, nationality, etc are gender agnostic. It is not a strong differentiator between records.

f. (1 point) Would blocking on gender be recommended for entity resolution. Explain.

No, because it leads to a very high number of false positives (low precision). Blocking on gender generates a large number of incorrect matches that can cause errors in the entity resolution process.

```
knitr::opts_chunk$set(echo = TRUE,
                      fig.width=4,
                      fig.height=3,
                      fig.align="center")
```