Probability and Bayes

(Conditional versus Marginal)

M Loecher

Elementary Definitions

Games of Chance

Conditional probabilities

Bayes Theorem

Exercises, Bayes

Naive Bayes

Appendix

Elementary Definitions

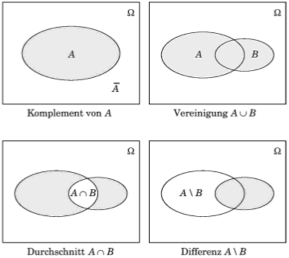
Union (“or”)

“A or B”

$$A \cup B$$

Intersection ("and")

$$A \cap B$$



Komplement von A

Vereinigung $A \cup B$

Durchschnitt $A \cap B$

Differenz $A \setminus B$

Multiplication rule for independent event

If two events A and B are independent:

$$P(A \cap B) = P(A) \cdot P(B)$$

Example 2 dice:

$$P(6,6) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

Complements

$$P(\bar{A}) = 1 - P(A)$$

Example 1 die:

$$P(w > 2) = 1 - P(w \leq 2)$$

Differences

Removing event B from A is written as $A \setminus B$. We have:

$$P(A \setminus B) = P(A) - P(A \cap B)$$

Example 1 die:

$$\text{P("even but no 2")} = \frac{3}{6} - \frac{1}{6}$$

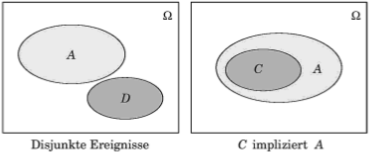
disjoint

Two events A and D are called “disjoint” if

$$A \cap B = \emptyset \Leftrightarrow P(A \cap B) = 0$$

Rule of addition for disjoint events:

$$P(A \cup B) = P(A) + P(B), \text{ if } A \cap B = \emptyset$$



Disjunkte Ereignisse

C impliziert *A*

BILDER 8.1 VENN-Diagramme zur Veranschaulichung von Ereignissen und Ereignisoperationen

Law of Addition I

For arbitrary events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

Example 1 die:

$$P(\text{even} \cup w \geq 4) =$$

Law of Addition II

If the events $A_1, \ldots, A_n$ are **pairwise disjoint**:

$$P(A_1 \cup A_2 \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

Example, 2 dice:

$$P(\text{"doubles"}) =$$

Fallacies

Kahnemann, Thinking, Fast and Slow

Linda: Less Is More

The best-known and most controversial of our experiments involved a fictitious lady called Linda. Amos and I made up the Linda problem to provide conclusive evidence of the role of heuristics in judgment and of their incompatibility with logic. This is how we described Linda:

Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

Which alternative is more probable?

- Linda is a bank teller.
- Linda is a bank teller and is active in the feminist movement

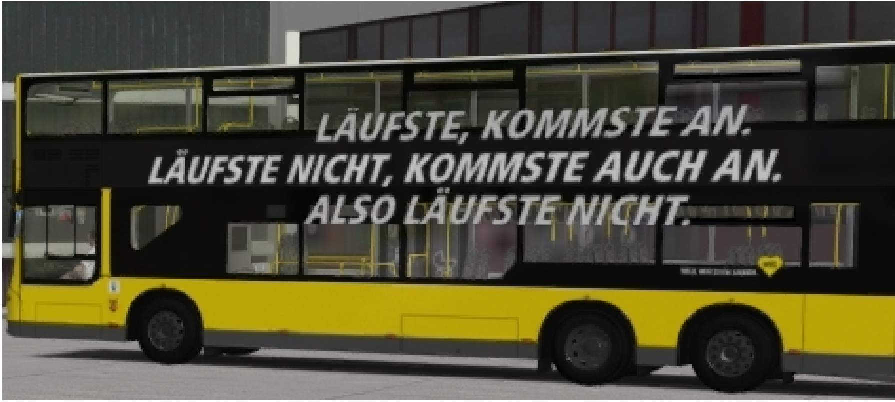
Games of Chance

Risk

1. "1 on 1" attack: P(A loses) =
2. "2 on 1" attack: P(A loses) =
3. "3 on 1" attack: P(A loses) =

Conditional probabilities

Even BVG knows them



Definition

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Easiest in terms of tables, here is the Titanic data set:

```
addmargins(table(titanic$Sex,titanic$Survived))
```

```
##
##             0   1 Sum
##   female   81 233 314
##   male    468 109 577
##   Sum     549 342 891
```

Independence, II

In general, this holds

$$P(A \cap B) = P(A) \cdot P(B|A)$$

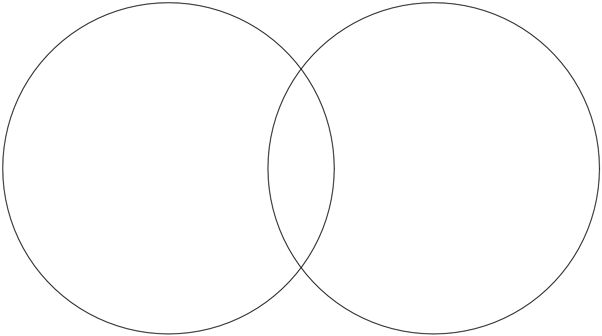
We can use this to define independence:

Two events A and B are **(stochastically) independent** if

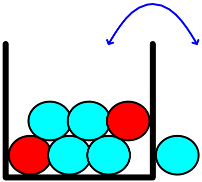
$$P(A|B) = P(A), \text{ or } P(B|A) = P(B)$$

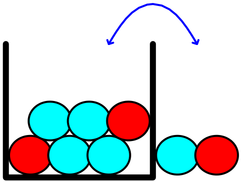
Venn Diagrams

$$P(\overline{A \cap B}) = P(\bar{A} \cup \bar{B})$$

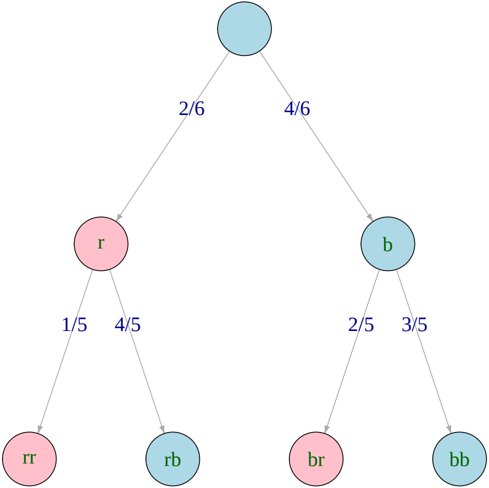


Urn

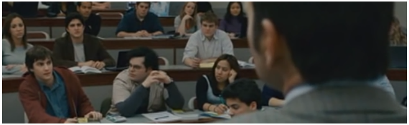




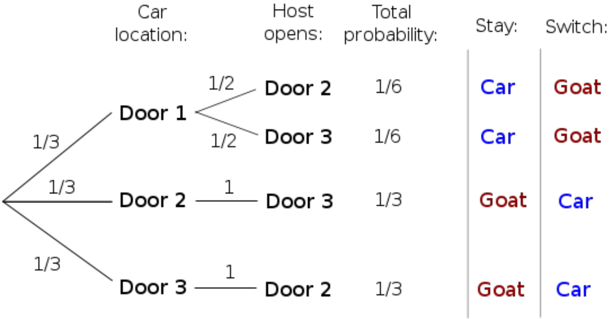
Tree diagrams



Monty Hall I



Monty Hall II



	Car location:		Host opens:	Total probability:	Stay:	Switch:
1/3	Door 1	1/2	Door 2	1/6	Car	Goat
		1/2	Door 3	1/6	Car	Goat
1/3	Door 2	1	Door 3	1/3	Goat	Car
1/3	Door 3	1	Door 2	1/3	Goat	Car

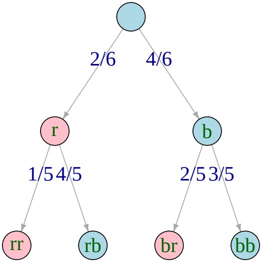
Total probability

$$\begin{aligned}
A &= (A \cap B) \cup (A \cap \bar{B}) \\
\Rightarrow P(A) &= P(A \cap B) + P(A \cap \bar{B}) \\
&= P(A) \cdot P(B|A) + P(A) \cdot P(\bar{B}|A) \\
&= P(B) \cdot P(A|B) + P(\bar{B}) \cdot P(A|\bar{B}) \\
&= \sum_k P(B_k) \cdot P(A|B_k)
\end{aligned}$$

and vice versa

$$\Rightarrow P(B) = \sum_k P(A_k) \cdot P(B|A_k)$$

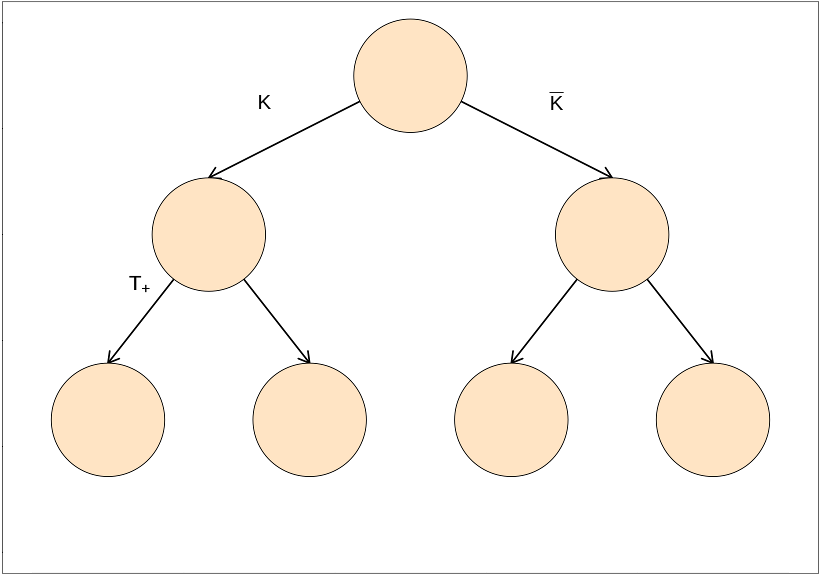
Total probability, Example I



$$
\begin{aligned}
P(\text{red ball on 2nd draw}) &= P(r_2) = P(r_2 \cap r_1) + P(r_2 \cap \bar{r_1}) \\
&= P(r_2 \cap r_1) + P(r_2 \cap b_1) \\
&= P(r_1) \cdot P(r_2|r_1) + P(\bar{r_1}) \cdot P(r_2|\bar{r_1}) \\
&= \frac{2}{6} \cdot \frac{1}{5} + \frac{4}{6} \cdot \frac{2}{5}
\end{aligned}
$$

Total probability, Example II

P(K) $= 0.01$ $P(T_+|K) = 0.98$ $P(T_-|\bar{K}) = 0.96$



Bayes Theorem

“Inverting” Conditional Probabilities

$$\begin{aligned}
P(A|B) &= \frac{P(A)\cdot P(B|A)}{P(B)} = \frac{P(A)}{P(B)}\cdot P(B|A) \\
&= \frac{P(A)\cdot P(B|A)}{P(B\cap A)+P(B\cap\bar{A})} \\
&= \frac{P(A)\cdot P(B|A)}{P(A)\cdot P(B|A)+P(\bar{A})\cdot P(B|\bar{A})} \\
&= \frac{P(A)\cdot P(B|A)}{\sum_k P(A_k)\cdot P(B|A_k)}
\end{aligned}$$

In the context observations/measurements do we call $P(A)$ the **prior** and $P(A|B)$ the **posterior probability**

Bayes Theorem, Example I

Kahnemann, Thinking, Fast and Slow

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. What is the probability that the cab involved in the accident was Blue rather than Green?

Bayes Theorem, Example II

Cliches/Stereotypes

- The proportions of green voters that commute to work by bike are similar in Holland and Deutschland. Are the inverse probabilities (i.e. how likely does someone vote for the green party once we know that she/he commutes to work by bike) similar as well ?
- Assume the proportion of truly right-wing extremists in the USA is about 5%. Let us further assume that 90% of those voted for Trump. How reasonable if the -often heard- "conclusion" that 90% of all Trump voters are right-wing extremists?

Bayes Theorem, Example III

Driving Gloves. “In 99% of all highway accidents the driver was not wearing gloves!”

Exercises, Bayes

Naive Bayes

Chain Rule for Probabilities

we can apply the definition of conditional probability to obtain:

$$P(A_n, \ldots, A_1) = \mathrm{P}(A_n | A_{n-1}, \ldots, A_1) \cdot \mathrm{P}(A_{n-1}, \ldots, A_1)$$

Repeating this process with each final term creates the product:

$$\mathrm{P}\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} \mathrm{P}\left(A_k \,\middle|\, \bigcap_{j=1}^{k-1} A_j\right)$$

With four variables, the chain rule produces this product of conditional probabilities:

$$\mathrm{P}(A_4, A_3, A_2, A_1) = \mathrm{P}(A_4 \mid A_3, A_2, A_1) \cdot \mathrm{P}(A_3 \mid A_2, A_1) \cdot \mathrm{P}(A_2 \mid A_1) \cdot \mathrm{P}(A_1)$$

Chain Rule, conditional version

$$\mathrm{P}\left(\bigcap_{k=1}^{n} A_k \mid D\right) = \prod_{k=1}^{n} \mathrm{P}\left(A_k \,\middle|\, \bigcap_{j=1}^{k-1} A_j, D\right) \tag{1}$$

With three variables plus a condition, we simply have:

$$\mathrm{P}(A_3, A_2, A_1 \mid D) = \mathrm{P}(A_3 \mid A_2, A_1, D) \cdot \mathrm{P}(A_2 \mid A_1, D) \cdot \mathrm{P}(A_1 \mid D)$$

Chain Rule, Exercises

- For the Titanic data verify

$$\mathrm{P}(F, P1 \mid S) = P(P1 \mid F, S) \cdot \mathrm{P}(F \mid S)$$

where “S = survived”, “F = female”, “P1 = Pclass 1”

- Compare with the simpler version

$$\mathrm{P}(P1 \mid S) \cdot \mathrm{P}(F \mid S)$$

Conditional Independence

In the standard notation of probability theory, A_1 and A_2 are **conditionally independent** given D if and only if the following (for **all** values of D) holds:

$$\mathrm{P}(A_2, A_1 \mid D) = \mathrm{P}(A_2 \mid D) \cdot \mathrm{P}(A_1 \mid D)$$

And for many variables:

$$\mathrm{P}\left(\bigcap_{k=1}^{n} A_k \mid D\right) = \prod_{k=1}^{n} \mathrm{P}\left(A_k \mid D\right)$$

Why do we care ?

Typical Classification asks

$$\mathrm{P}\left(D \mid \bigcap_{k=1}^{n} A_k\right)$$

Example: The *HouseVotes84* dataset describes how 435 representatives voted - yes (y), no (n) or unknown (NA) - on 16 key issues presented to Congress. The dataset also provides the party affiliation of each representative - democrat or republican.

```r
data(HouseVotes84, package = "mlbench")
source("Impute_NAs.R")
```

Naturally, we would like to figure out the party affiliation from a knowledge of voting patterns

HouseVotes, cont.

By Bayes theorem, this goal can be recast as

$$\mathrm{P}\left(D \mid \bigcap_{k=1}^{n} A_k\right) = P\left(\bigcap_{k=1}^{n} A_k \mid D\right) \cdot \frac{P(D)}{P\left(\bigcap_{k=1}^{n} A_k\right)}$$

which can be **significantly** simplified if we make the conditional independence assumption:

$$\mathrm{P}\left(D \mid \bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} \mathrm{P}\left(A_k \mid D\right) \cdot \frac{P(D)}{P\left(\bigcap_{k=1}^{n} A_k\right)} \tag{2}$$

HouseVotes, cont.

The assumption of independent conditional probabilities is a drastic one. What it is saying is that the features are completely independent of each other, given a party affiliation. This is clearly not the case in the situation above: how representatives vote on a particular issue is coloured by their beliefs and values. However, the Naive Bayes assumption works surprisingly well for this problem as it does in many other situations where we know upfront that it is grossly incorrect.

Performance

```
##              true
## pred         democrat republican
##   democrat         55          3
##   republican        8         43
```

```
## [1] 0.8990826
```

“Dropping the denominator”

IF we only care about the ranking of the probabilities, not their exact values, do we need the denominator in Eq (2) ?

Exercises, Naive Bayes

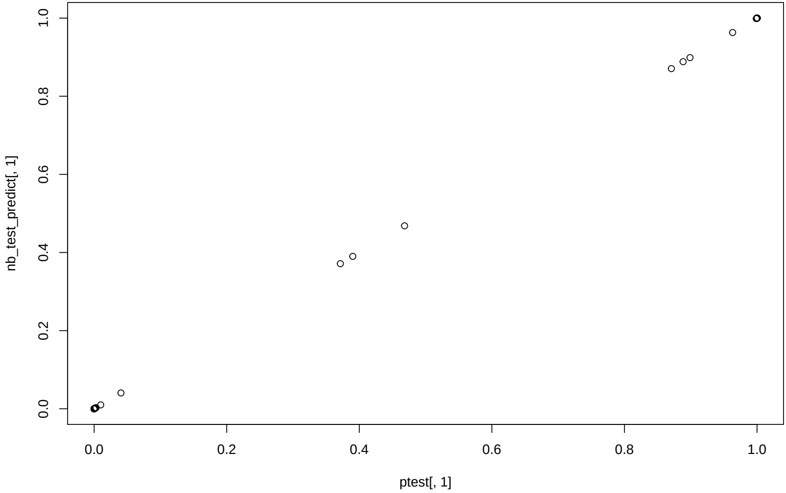
1. House Votes:

- Verify the conditional chain rule Eq (1)
- Compute the most likely class “manually”, using Eq (2), for the test data and compare with the package.

2. Movie Reviews

- Follow the steps in https://cran.r-project.org/web/packages/text2vec/vignettes/text-vectorization.html
- Compare a Naive Bayes model with the glmnet model proposed.

Solutions



Low Counts

Let's say you've trained your Naive Bayes Classifier on 2 classes, "Ham" and "Spam" (i.e. it classifies emails).

Now let's say you have an email $(w_1, w_2, ..., w_n)$ which your classifier rates very highly as "Ham", say

$$P(w_1, w_2, ...w_n|Ham) = .90,\ P(w_1, w_2, ..w_n|Spam) = .10$$

Now let's say you have another email $(w_1, w_2, ..., w_n, w_{n+1})$ which is exactly the same as the above email except that there's one word in it that isn't included in the vocabulary. Therefore, since this word's count is 0, $P(w_{n+1}|Ham) = P(w_{n+1}|Spam) = 0$

Suddenly,

$$P(w_1, w_2, ...w_n, w_{n+1}|Ham) = P(w_1, w_2, ...w_n|Ham) * P(w_{n+1}|Ham) = 0$$

and

$$P(w_1, w_2, ..w_n, w_{n+1}|Spam) = P(w_1, w_2, ...w_n|Spam) * P(w_{n+1}|Spam) = 0$$

Despite the 1st email being strongly classified in one class, this 2nd email may be classified differently because of that last word having a probability of zero.

Laplace smoothing

$$P(X_i = x_i) = \frac{\text{count}(X_i = x_i)}{N}$$

Laplace smoothing solves this by giving the last word a small non-zero probability for both classes, so that the posterior probabilities don't suddenly drop to zero. "Pretend you've seen each variable k extra times"

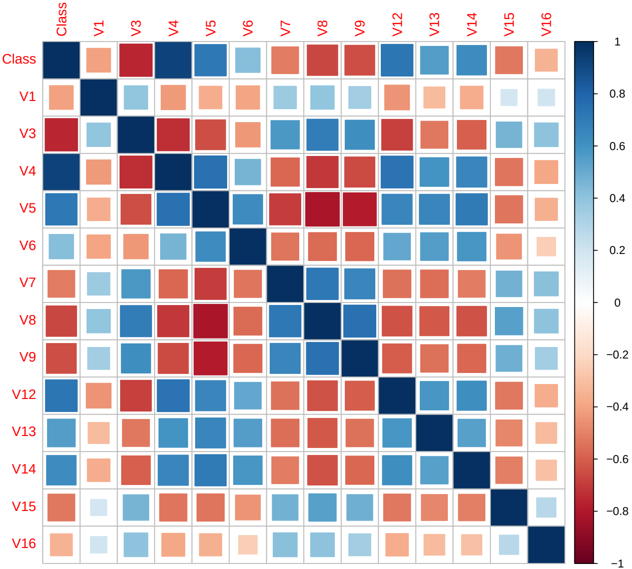
$$P(X_i = x_i) = \frac{\text{count}(X_i = x_i) + k}{N + k\#(X_i)}$$

where $\#(X_i)$ is the number of values X_i can assume.

https://classes.soe.ucsc.edu/cmps140/Winter17/slides/3.pdf

Appendix

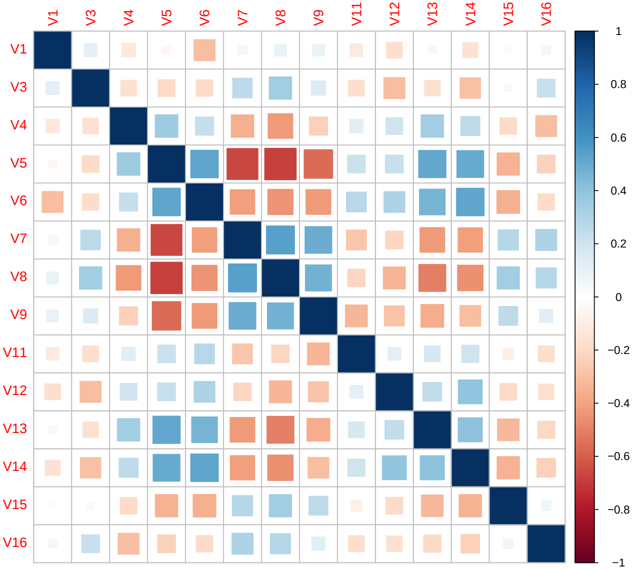
Correlations



Definitely not independent !!

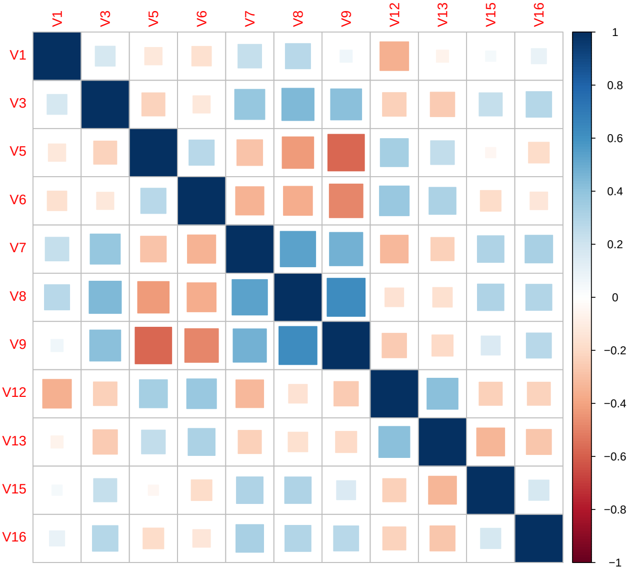
Conditional Independence I

Conditional on Democrat



Conditional Independence II

Conditional on Republican



Chain rule and trees

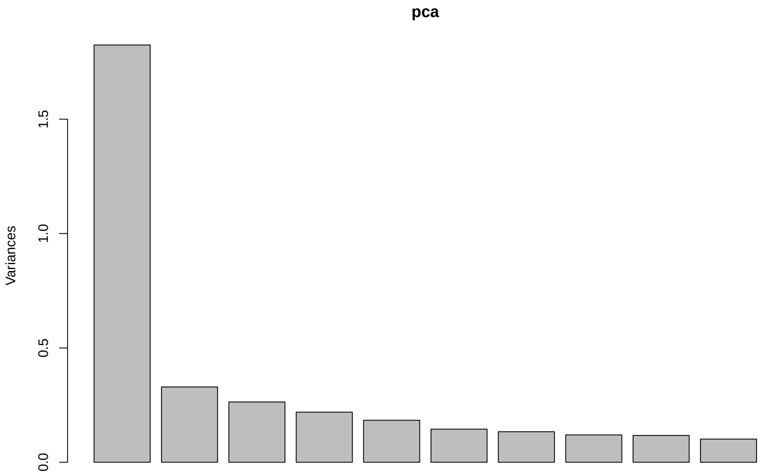
Sparsity

Curse of Dimensionality

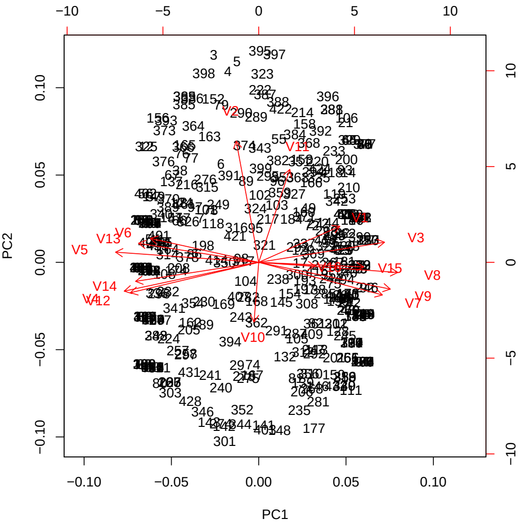
```r
nrow(unique(HouseVotes84))/2^(17)
```

```
## [1] 0.002090454
```

PCA



PCA II


PC1
PC2
−10
−5
0
5
10
−0.10
−0.05
0.00
0.05
0.10
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15