Density Estimation Demo

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Outline

Credit Risk Data

Recall the credit data:

```
b,30.83,0,u,g,w,v,1.25,t,t,01,f,g,00202,0,+
a,58.67,4.46,u,g,q,h,3.04,t,t,06,f,g,00043,560,+
a,24.50,0.5,u,g,q,h,1.5,t,f,0,f,g,00280,824,+
b,27.83,1.54,u,g,w,v,3.75,t,t,05,t,g,00100,3,+
b,20.17,5.625,u,g,w,v,1.71,t,f,0,f,s,00120,0,+
b,32.08,4,u,g,m,v,2.5,t,f,0,t,g,00360,0,+
b,33.17,1.04,u,g,r,h,6.5,t,f,0,t,g,00164,31285,+
a,22.92,11.585,u,g,cc,v,0.04,t,f,0,f,g,00080,1349,+
b,54.42,0.5,y,p,k,h,3.96,t,f,0,f,g,00180,314,+
b,42.50,4.915,y,p,w,v,3.165,t,f,0,t,g,00052,1442,+
b,22.08,0.83,u,g,c,h,2.165,f,f,0,t,g,00128,0,+
b,29.92,1.835,u,g,c,h,4.335,t,f,0,f,g,00260,200,+
a,38.25,6,u,g,k,v,1,t,f,0,t,g,00000,0,+
b,48.08,6.04,u,g,k,v,0.04,f,f,0,f,g,00000,2690,+
a,45.83,10.5,u,g,q,v,5,t,t,07,t,g,00000,0,+
```

Credit Risk Data

It has both continuous and categorical variables, which are things like:

- Status of checking account (low balance, medium balance, high balance, no account)
- Credit history (no credits, all credits paid back, past delays in paying, ...)
- Reason for requesting credit (purchase car, repair house, ...)
- Credit amount requested
- Marital status
- Age
- Whether the person has been rated as a good $(+)$ or bad $(-)$ credit risk.

Credit Data

- When we previously predicted credit risk using naive Bayes, we discretized all the continuous predictors. Today we will do some naive Bayes predictions, using the continuous predictors without discretization.

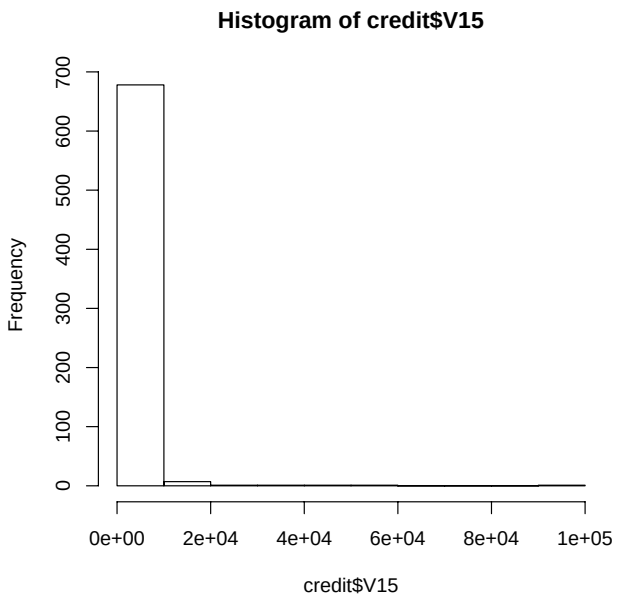
- Download the credit data from Blackboard (both files). The data are in “crx.data” and information about the dataset (metadata) is in “crx.names”

- Read the dataset into R:

  credit = read.table(“C:/temp/crx.data”, sep = “,”, na.strings = “?”)

Credit Data

Create a histogram of V15:



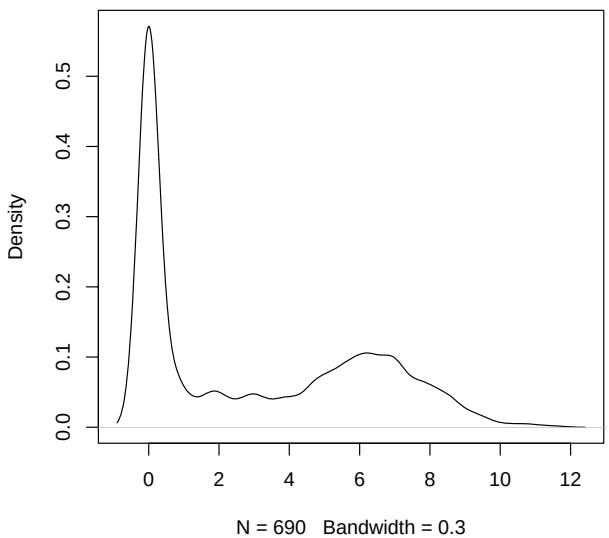
It is a strongly right-skewed variable, so take a log transformation.

Credit Data Analysis

Now do a kernel density estimate of the new variable, picking an arbitrary bandwidth (“bw”):

dens = density(x =credit$logV15, na.rm = T, bw = .3)
plot(dens)

What does “na.rm = T” do here?



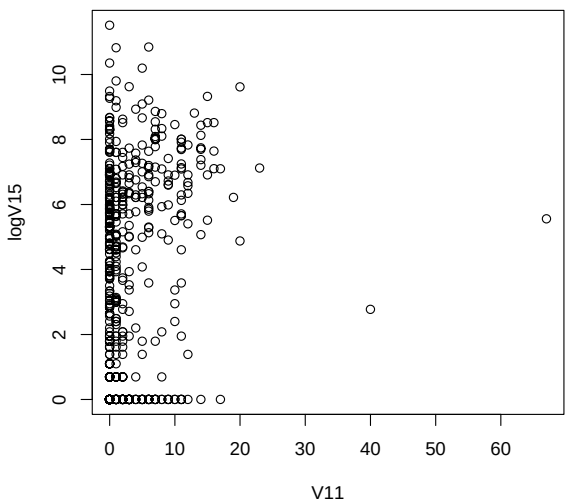
Credit Data Analysis

This choice of bandwidth appears to be too small, because the density estimate looks wiggly. Try using the default choice of bandwidth instead, and re-plot. Better?

dens = density(x=credit$logV15, na.rm = T)

This uses a rule-of-thumb choice of the bandwidth that is a function of the standard deviation of the variable, and the sample size (for larger sample sizes, a smaller bandwidth can be used)

Let's create a 2-dimensional density estimate for the joint density of V11 and logV15. First, create a scatterplot:



Credit Data Analysis

- Install the MASS package by going to Packages->Install Packages in R. Select a “mirror” site in the U.S., then select MASS from the list. Load the package:
  library(MASS)

- We will use the “kde2d” function, but it cannot handle missing data. Apply “kde2d” to the non-missing values of V11 and logV15:
  miss = is.na(credit$V11) | is.na(credit$logV15)
  dens = kde2d(x = credit$V11[!miss], y = credit$logV15[!miss], n = 200, h = c(10, 1.4))

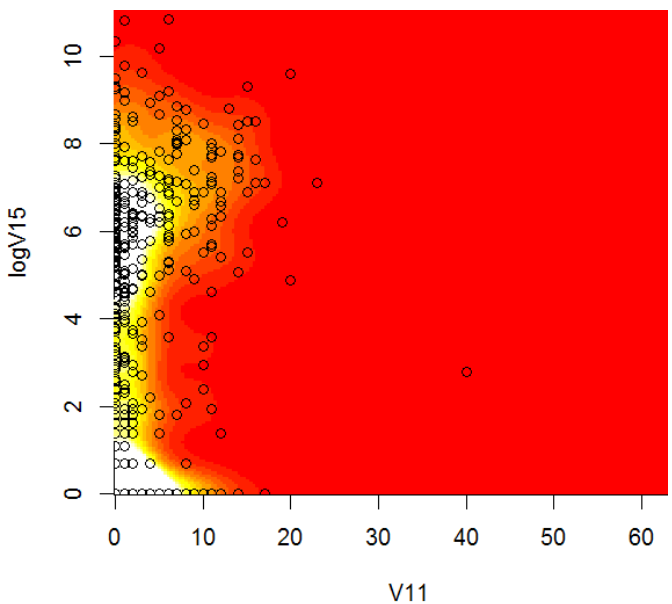
- Here “n” is specifying the size of the grid on which we want to evaluate the density estimate; this will affect how our plot looks.

- “h” is the bandwidth; the default choices do not work well in this case so I tried a number of different values and these work well.

Credit Data Analysis

Create a “heat plot” for the two-dimensional density estimate. This is a way of displaying a function of two variables; dark red means the function is low, while light yellow or white means the function is very high. Code:

Overlay the observed values (scatterplot) on the heat map:
points(x = credit$V11[!miss], y = credit$logV15[!miss])



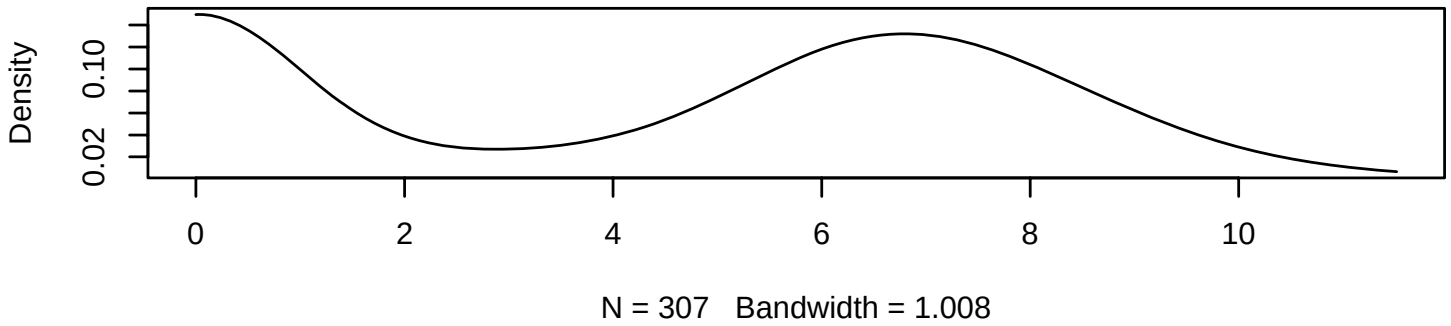
Credit Data Analysis

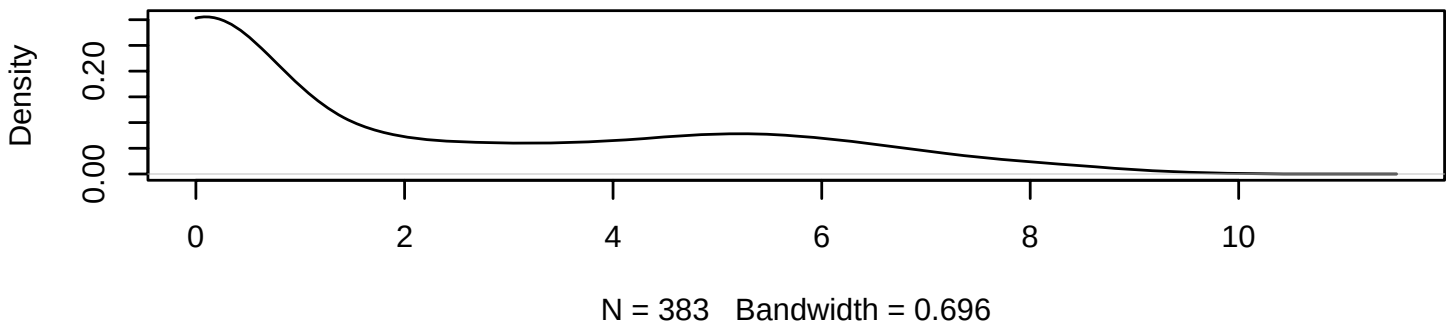
To apply naive Bayes with the single predictor variable $X_1 = \text{logV15}$, we have to estimate the density functions $f_{X_1|Y=+}(z)$ and $f_{X_1|Y=-}(z)$ for $z \in \mathbb{R}$. Here's some code to get and plot them:

```
densPlusV15 = density( credit$logV15[credit$V16==”+”], na.rm = T,
from = min(credit$logV15), to = max(credit$logV15), n=512 )
densMinusV15 = density( credit$logV15[credit$V16==”-”], na.rm = T,
from = min(credit$logV15), to = max(credit$logV15), n=512 )
```

```
par(mfrow = c(3,1)
plot(densPlusV15, main = “”)
plot(densMinusV15, main = “”)
```

Credit Data Analysis





Credit Data Analysis

Now calculate the predicted probability that $Y = +$, given that $X_1 = x_1$ as a function of x_1, which is given by
$\Pr(Y = +|X_1 = x_1) = \frac{\Pr(Y=+)f_{X_1|Y=+}(x_1)}{\Pr(Y=+)f_{X_1|Y=+}(x_1)+\Pr(Y=-)f_{X_1|Y=-}(x_1)}$. Code:

```
probPlus = mean(credit$V16 == "+")
probGivenV15 = probPlus * densPlusV15$y /
(probPlus * densPlusV15$y + (1-probPlus) * densMinusV15$y)
plot( x = densPlusV15$x, y = probGivenV15, type = "l", ylim = c(0,1))
```

Credit Data Analysis