Graphical Models

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Outline

Credit Risk Data

Dataset on individuals applying for credit (from Quinlan 1986):

```
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,+
```

The last column indicates whether the person has been rated as a good ($+$) or bad ($-$) credit risk.

Credit Risk Data

There are both continuous and discrete predictors (categories coded as letters)

These predictors 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

Credit Data Analysis

- On Wednesday we'll apply naive Bayes to this dataset in class.
- We do not yet know how to fit naive Bayes using continuous predictors (will learn later in course)
- So on Wednesday we'll discretize each continuous predictor, i.e. turn it into a discrete random variable. One simple approach is to split at the median, creating one value if the predictor is less than its median, and another value if it is greater than its median.

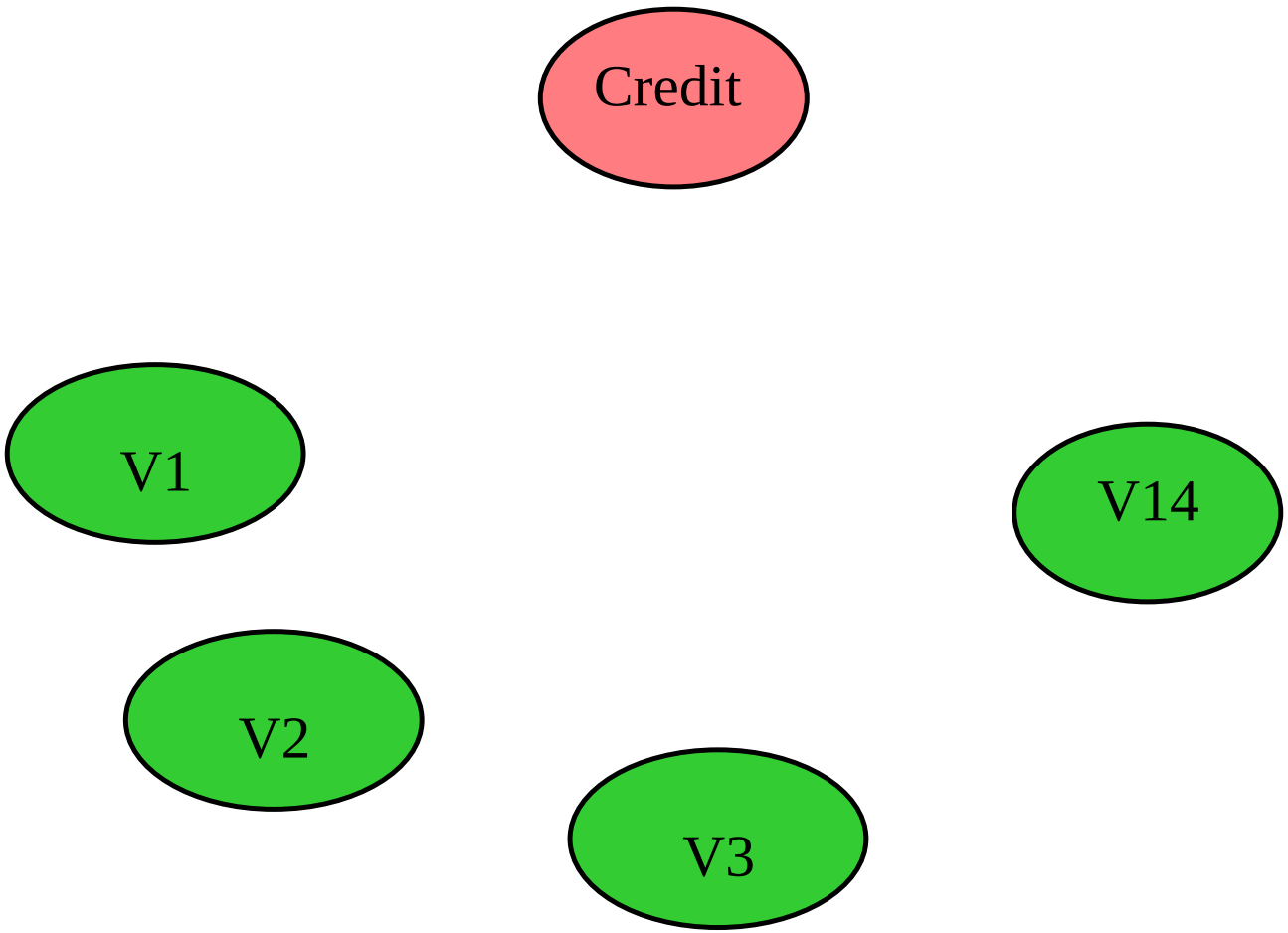
Naive Bayes Assumption:

$$\Pr(\{X_1, X_2, \ldots, X_K\}|Y) = \prod_{k=1}^{K} \Pr(X_k|Y)$$

- Naive Bayes assumes that, conditional on the outcome variable, the predictors are independent.

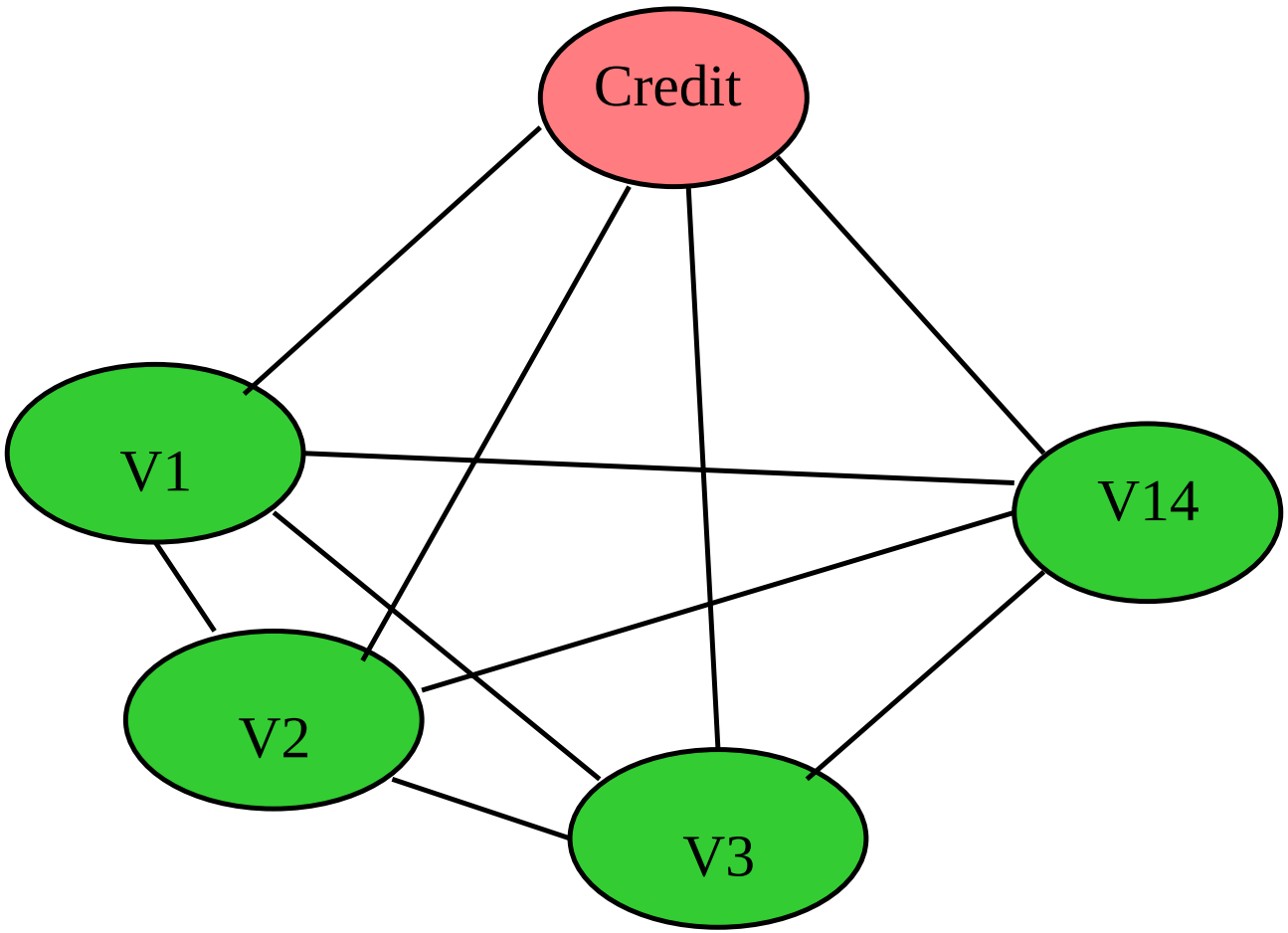
Credit Variables

Recall the variables for the credit data:



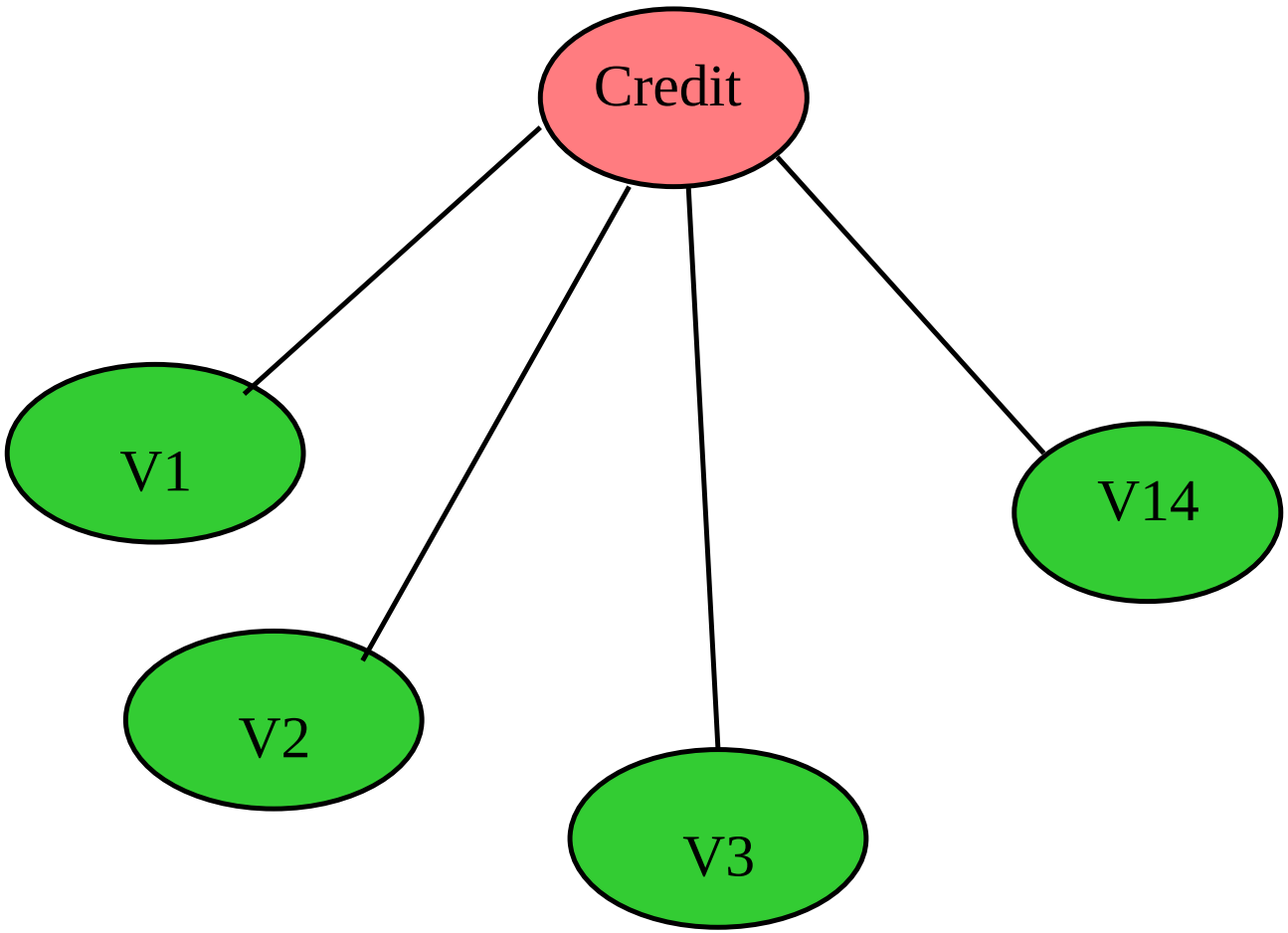
A Full Model

A full model for $\Pr(Y, \{X_k\})$ must take into account all the two-way, three-way, etc. interactions between the variables:



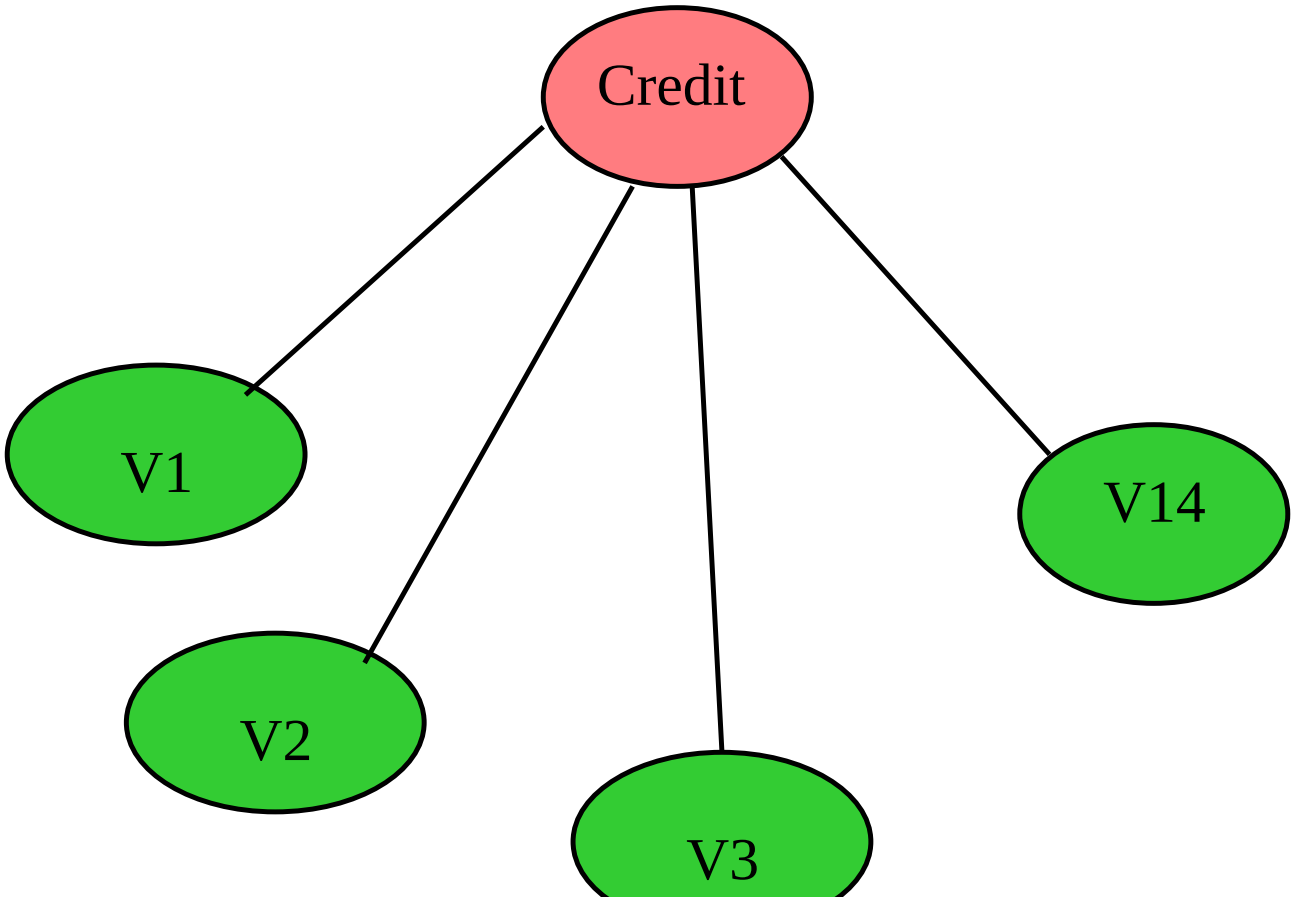
Naive Bayes

Naive Bayes only models the interaction between the outcome and each of the predictors:



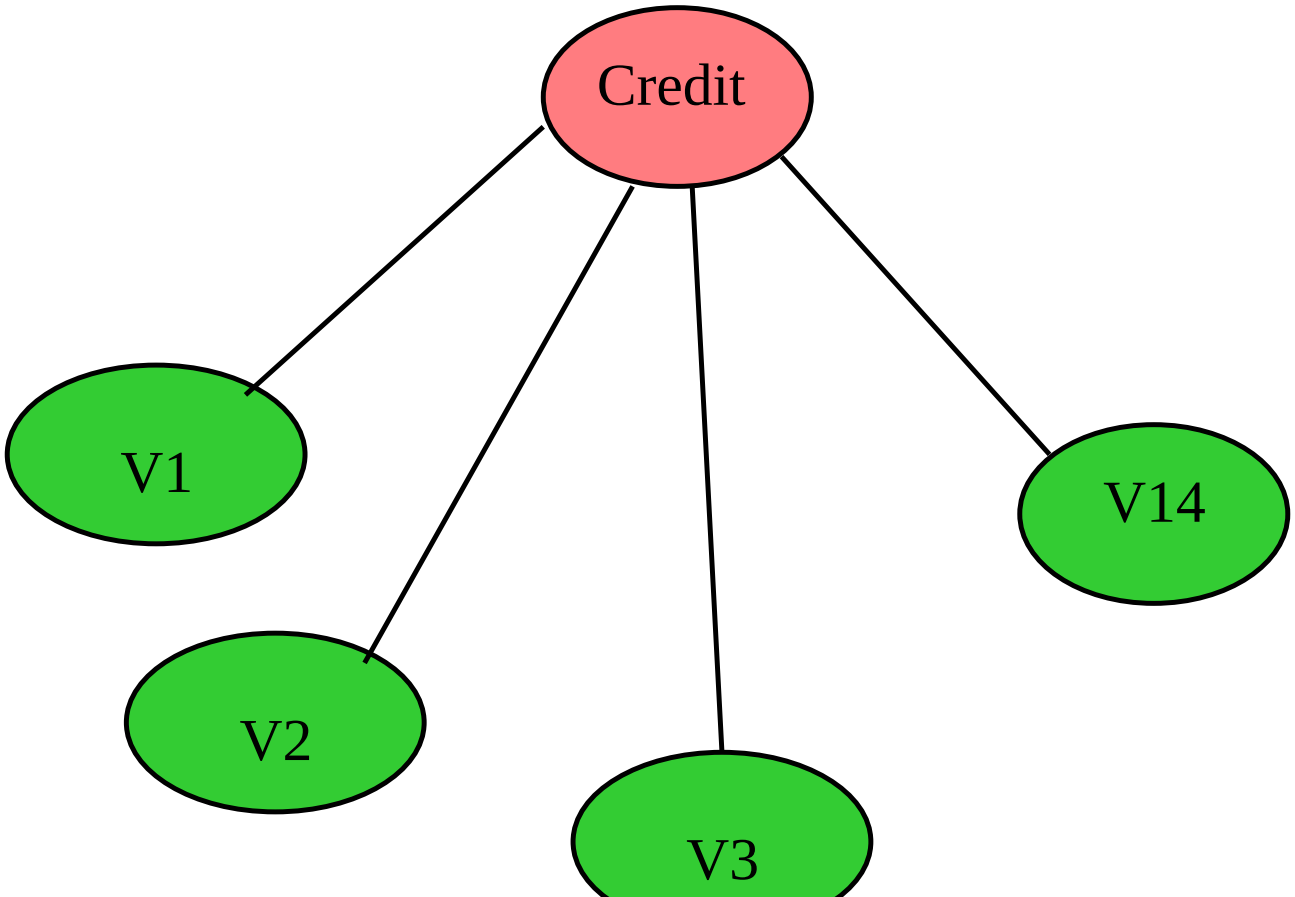
Naive Bayes

This type of graph is called a conditional independence graph or undirected graphical model and shows the model's conditional independence assumptions:



Naive Bayes

For instance:



Credit Data

- Do you think the naive Bayes assumption holds (approximately) for the credit data?
- Recall that the predictors are variables like:
  - Amount in checking account
  - Duration of credit in months
  - Credit history
  - Duration of present employment
  - Marital status

Credit Data

- Let's find the most dependent pairs of predictors, conditional on Y.
- For the pair V1 and V2 (for instance), look at the joint counts table for observations in the training data having $Y = +$:

	V2 = FALSE	V2 = TRUE
V1 = a	28	30
V1 = b	60	76

and for observations having $Y = -$:

	V2 = FALSE	V2 = TRUE
V1 = a	47	28
V1 = b	86	87

Credit Data

- Are these two variables highly dependent, conditional on Y?
- Divide the joint counts table entries by the row sums to get the conditional frequencies $\hat{Pr}(\text{V2}|\text{V1}, Y = +)$:

	V2 = FALSE	V2 = TRUE
V1 = a	0.48	0.52
V1 = b	0.44	0.56

and $\hat{Pr}(\text{V2}|\text{V1}, Y = -)$:

	V2 = FALSE	V2 = TRUE
V1 = a	0.63	0.37
V1 = b	0.50	0.50

Credit Data

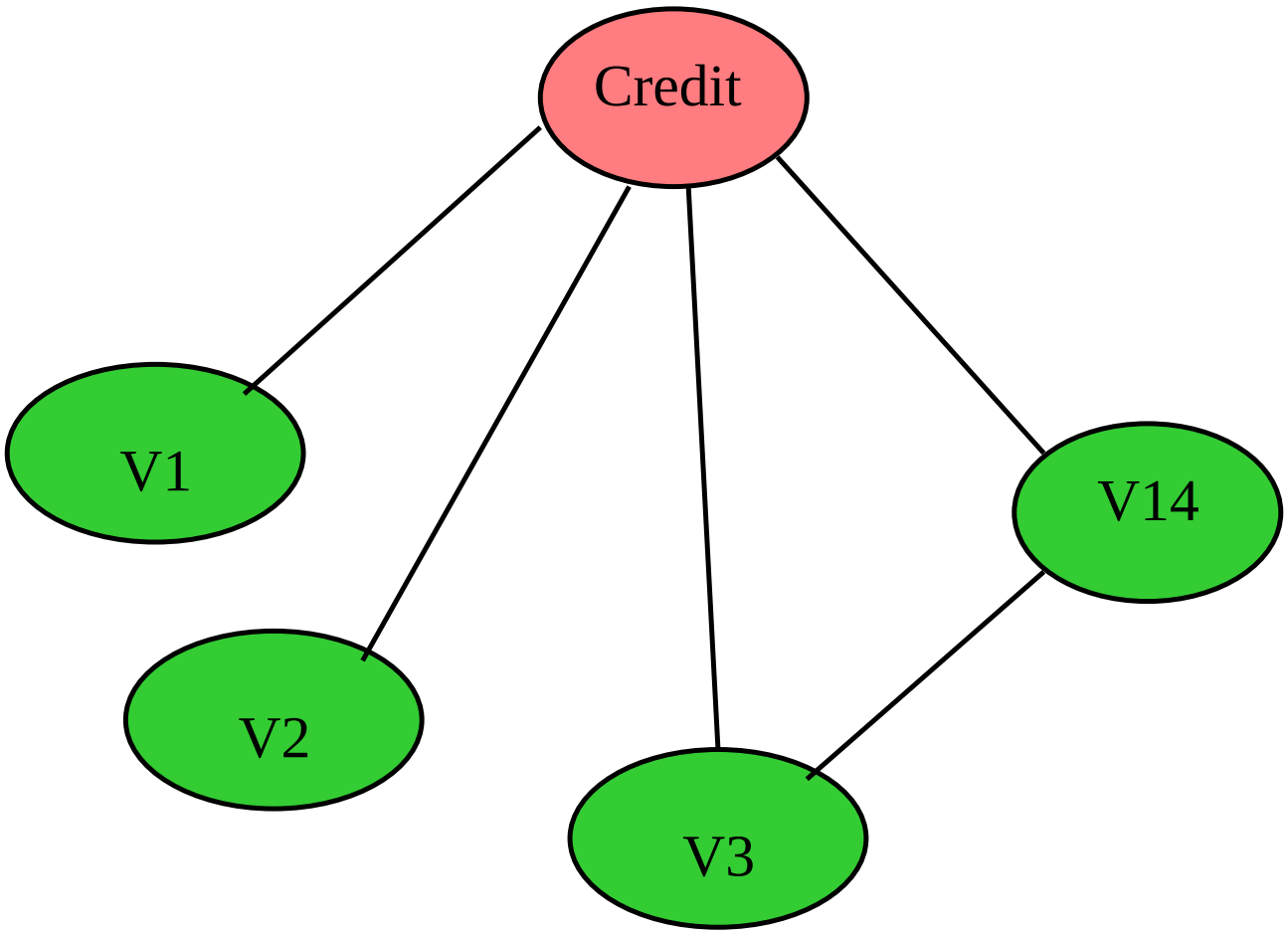
- The two variables V3 and V14 are strongly dependent. Here is $\hat{Pr}(\text{V3}|\text{V14}, Y = +)$:

	V3 = FALSE	V3 = TRUE
V14 = FALSE	0.29	0.71
V14 = TRUE	0.56	0.44

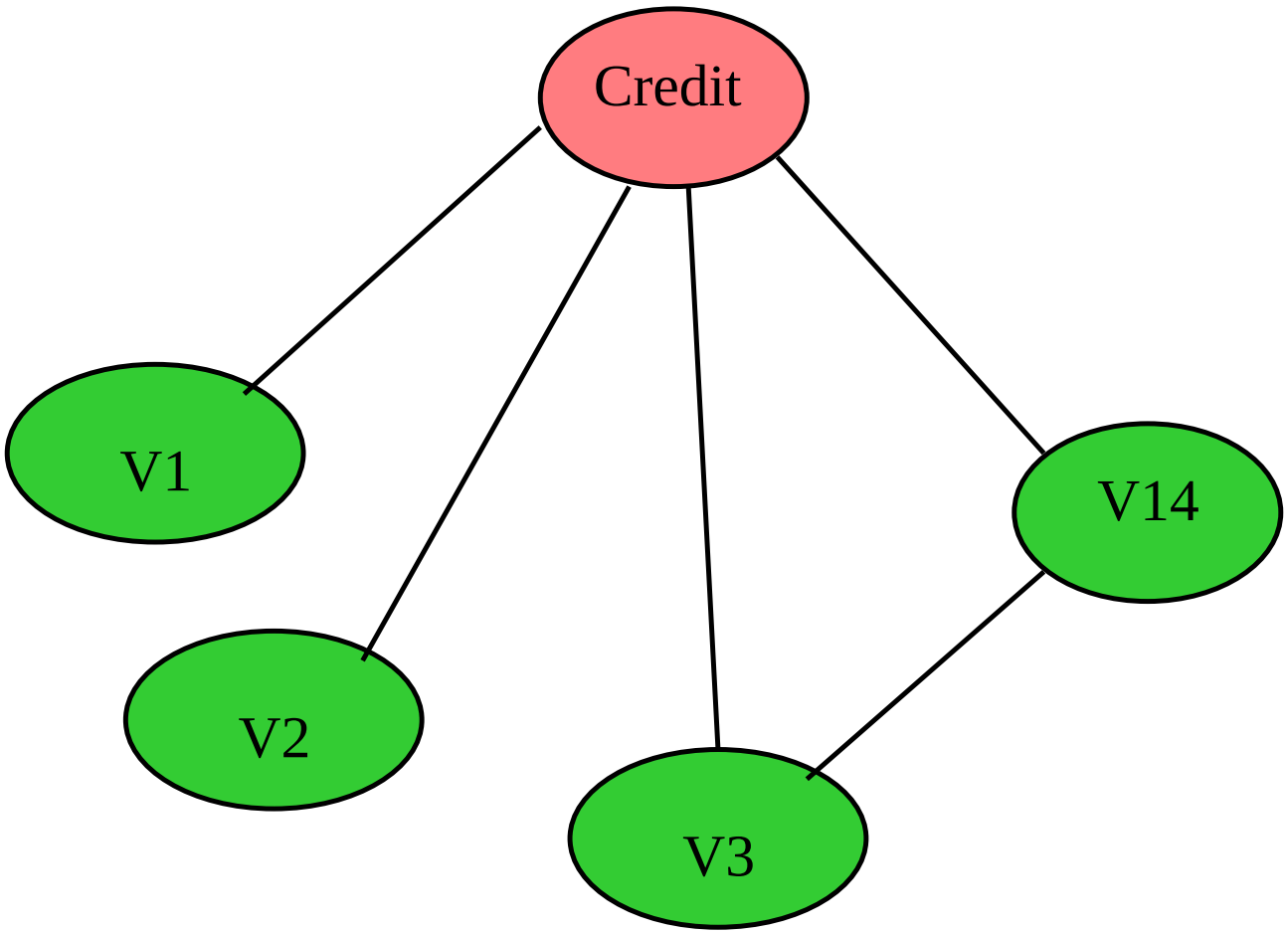
and $\hat{Pr}(\text{V3}|\text{V14}, Y = -)$:

	V3 = FALSE	V3 = TRUE
V14 = FALSE	0.53	0.47
V14 = TRUE	0.66	0.34

We could relax the naive Bayes assumption by adding an edge between V3 and V14:



I.e.:



Graphical Models

How does adding an edge between two predictors X_i and X_j change the naive Bayes calculations?

Graphical Models

- All we have to do in R to fit and predict for our graphical model is:
  - Replace the V3 and V14 variables in the credit data set with a single variable that encodes both V3 and V14 together
  - I.e., this variable has one value corresponding to each combination of values for V3 and V14
  - Call the naive Bayes functions (nb.train and nb.predict) on the new credit data set

Graphical Models

- Here was the original joint counts table for V3 and V14:

	V3 = FALSE	V3 = TRUE
V14 = FALSE	92	141
V14 = TRUE	134	84

- Here is the counts vector for the recoded variable:

1	2	3	4
92	134	141	84

Graphical Models

- Adding the edge resulted in a slight decrease in the overall error rate on the test data, from 14.7% to 14.3%
- The test data set was small (230), so this was only one less misclassification
- To really see clearly the effect, we would need a larger test data set
- You will create a graphical model in lab for a much larger data set!

Graphical Models

- Is it always better to add more edges?

Graphical Models

- Pretend we have 4 predictors and all of them are binary (0-1)
- For the naive Bayes method, how many parameters are there that we estimate in the training step?

  A. 1-4

  B. 5-8

  C. 9-12

  D. 13-16

Graphical Models

- If we draw an edge between predictors 1 and 2 to create a graphical model, how many parameters do we need to estimate in the training step?
- If we then add edges from predictors 1 and 2 to predictor 3, how many parameters are there?
- How about if all of the predictors have edges in between?