Fundamentals of Machine Learning

Chapter 6: Probability-based Learning

Sections 6.4, 6.5

Smoothing

Example: Loan application fraud detection

ID	Credit History	Guarantor/ CoApplicant	Accommodation	Fraud
1	current	none	own	true
2	paid	none	own	false
3	paid	none	own	false
4	paid	guarantor	rent	true
5	arrears	none	own	false
6	arrears	none	own	true
7	current	none	own	false
8	arrears	none	own	false
9	current	none	rent	false
10	none	none	own	true
11	current	coapplicant	own	false
12	current	none	own	true
13	current	none	rent	true
14	paid	none	own	false
15	arrears	none	own	false
16	current	none	own	false
17	arrears	coapplicant	rent	false
18	arrears	none	free	false
19	arrears	none	own	false
20	paid	none	own	false

$$
\begin{array}{rclrcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(\text{CH} = \textit{'none'} \mid fr) & = & 0.1666 & P(\text{CH} = \textit{'none'} \mid \neg fr) & = & 0 \\
P(\text{CH} = \textit{'paid'} \mid fr) & = & 0.1666 & P(\text{CH} = \textit{'paid'} \mid \neg fr) & = & 0.2857 \\
P(\text{CH} = \textit{'current'} \mid fr) & = & 0.5 & P(\text{CH} = \textit{'current'} \mid \neg fr) & = & 0.2857 \\
P(\text{CH} = \textit{'arrears'} \mid fr) & = & 0.1666 & P(\text{CH} = \textit{'arrears'} \mid \neg fr) & = & 0.4286 \\
P(\text{GC} = \textit{'none'} \mid fr) & = & 0.8334 & P(\text{GC} = \textit{'none'} \mid \neg fr) & = & 0.8571 \\
P(\text{GC} = \textit{'guarantor'} \mid fr) & = & 0.1666 & P(\text{GC} = \textit{'guarantor'} \mid \neg fr) & = & 0 \\
P(\text{GC} = \textit{'coapplicant'} \mid fr) & = & 0 & P(\text{GC} = \textit{'coapplicant'} \mid \neg fr) & = & 0.1429 \\
P(\text{ACC} = \textit{'own'} \mid fr) & = & 0.6666 & P(\text{ACC} = \textit{'own'} \mid \neg fr) & = & 0.7857 \\
P(\text{ACC} = \textit{'rent'} \mid fr) & = & 0.3333 & P(\text{ACC} = \textit{'rent'} \mid \neg fr) & = & 0.1429 \\
P(\text{ACC} = \textit{'free'} \mid fr) & = & 0 & P(\text{ACC} = \textit{'free'} \mid \neg fr) & = & 0.0714
\end{array}
$$

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMMODATION	FRAUDULENT
paid	guarantor	free	?

$$
\begin{array}{rclrcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(CH = paid \mid fr) & = & 0.1666 & P(CH = paid \mid \neg fr) & = & 0.2857 \\
P(GC = guarantor \mid fr) & = & 0.1666 & P(GC = guarantor \mid \neg fr) & = & 0 \\
P(ACC = free \mid fr) & = & 0 & P(ACC = free \mid \neg fr) & = & 0.0714
\end{array}
$$

$$
\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.0
$$

$$
\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.0
$$

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMMODATION	FRAUDULENT
paid	guarantor	free	?

- The standard way to avoid this issue is to use **smoothing**.
- Smoothing takes some of the probability from the events with lots of the probability share and gives it to the other probabilities in the set.

- There are several different ways to smooth probabilities, we will use **Laplacian smoothing**.

Laplacian Smoothing (conditional probabilities)

$$P(f = v|t) = \frac{count(f = v|t) + k}{count(f|t) + (k \times |Domain(f)|)}$$

Raw	$P(GC = none \mid \neg fr)$	$=$	0.8571
Probabilities	$P(GC = guarantor \mid \neg fr)$	$=$	0
	$P(GC = coapplicant \mid \neg fr)$	$=$	0.1429
Smoothing	k	$=$	3
Parameters	$count(GC \mid \neg fr)$	$=$	14
	$count(GC = none \mid \neg fr)$	$=$	12
	$count(GC = guarantor \mid \neg fr)$	$=$	0
	$count(GC = coapplicant \mid \neg fr)$	$=$	2
	$\lvert Domain(GC) \rvert$	$=$	3
Smoothed	$P(GC = none \mid \neg fr) = \frac{12+3}{14+(3\times3)}$	$=$	0.6522
Probabilities	$P(GC = guarantor \mid \neg fr) = \frac{0+3}{14+(3\times3)}$	$=$	0.1304
	$P(GC = coapplicant \mid \neg fr) = \frac{2+3}{14+(3\times3)}$	$=$	0.2174

Table: Smoothing the posterior probabilities for the GUARANTOR/COAPPLICANT feature conditioned on FRAUDULENT being False.

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = none \mid fr)$	$=$	0.2222	$P(CH = none \mid \neg fr)$	$=$	0.1154
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(CH = current \mid fr)$	$=$	0.3333	$P(CH = current \mid \neg fr)$	$=$	0.2692
$P(CH = arrears \mid fr)$	$=$	0.2222	$P(CH = arrears \mid \neg fr)$	$=$	0.3462
$P(GC = none \mid fr)$	$=$	0.5333	$P(GC = none \mid \neg fr)$	$=$	0.6522
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(GC = coapplicant \mid fr)$	$=$	0.2	$P(GC = coapplicant \mid \neg fr)$	$=$	0.2174
$P(ACC = own \mid fr)$	$=$	0.4667	$P(ACC = own \mid \neg fr)$	$=$	0.6087
$P(ACC = rent \mid fr)$	$=$	0.3333	$P(ACC = rent \mid \neg fr)$	$=$	0.2174
$P(ACC = Free \mid fr)$	$=$	0.2	$P(ACC = Free \mid \neg fr)$	$=$	0.1739

Table: The Laplacian smoothed, with $k = 3$, probabilities needed by a Naive Bayes prediction model calculated from the fraud detection dataset. Notation key: FR=FRAUDULENT, CH=CREDIT HISTORY, GC = GUARANTOR/COAPPLICANT, ACC = ACCOMODATION, T=*'True'*, F=*'False'*.

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMMODATION	FRAUDULENT
paid	guarantor	free	?

$$
\begin{array}{rcl rcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(CH = paid|fr) & = & 0.2222 & P(CH = paid|\neg fr) & = & 0.2692 \\
P(GC = guarantor|fr) & = & 0.2667 & P(GC = guarantor|\neg fr) & = & 0.1304 \\
P(ACC = Free|fr) & = & 0.2 & P(ACC = Free|\neg fr) & = & 0.1739
\end{array}
$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[m]|fr)\right) \times P(fr) = 0.0036$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[m]|\neg fr)\right) \times P(\neg fr) = 0.0043$$

Table: The relevant smoothed probabilities, from Table 2 [9], needed by the Naive Bayes prediction model in order to classify the query from the previous slide and the calculation of the scores for each candidate classification.

Continuous Features: Probability Density Functions

Table: The dataset from the loan application fraud detection domain with a new continuous descriptive features added: ACCOUNT BALANCE

ID	CREDIT HISTORY	GUARANTOR/ COAPPLICANT	ACCOMMODATION	ACCOUNT BALANCE	FRAUD
1	current	none	own	56.75	true
2	current	none	own	1,800.11	false
3	current	none	own	1,341.03	false
4	paid	guarantor	rent	749.50	true
5	arrears	none	own	1,150.00	false
6	arrears	none	own	928.30	true
7	current	none	own	250.90	false
8	arrears	none	own	806.15	false
9	current	none	rent	1,209.02	false
10	none	none	own	405.72	true
11	current	coapplicant	own	550.00	false
12	current	none	free	223.89	true
13	current	none	rent	103.23	true
14	paid	none	own	758.22	false
15	arrears	none	own	430.79	false
16	current	none	own	675.11	false
17	arrears	coapplicant	rent	1,657.20	false
18	arrears	none	free	1,405.18	false
19	arrears	none	own	760.51	false
20	current	none	own	985.41	false

(PDF: Probability Density Function)

- We need to define two PDFs for the new ACCOUNT BALANCE (AB) feature with each PDF conditioned on a different value in the domain or the target:
  - $P(AB = X|fr) = PDF_1(AB = X|fr)$
  - $P(AB = X|\neg fr) = PDF_2(AB = X|\neg fr)$
- Note that these two PDFs do not have to be defined using the same statistical distribution.

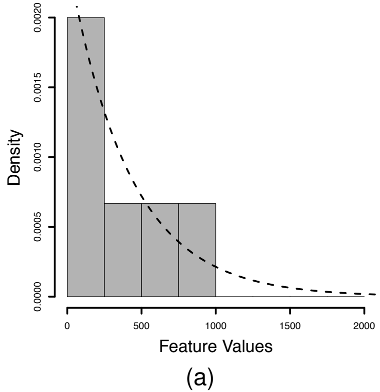

Figure: Histograms, using a bin size of 250 units, and density curves for the ACCOUNT BALANCE feature: (a) the fraudulent instances overlaid with a fitted exponential distribution; (b) the non-fraudulent instances overlaid with a fitted normal distribution.

- From the shape of these histograms it appears that
  - the distribution of values taken by the ACCOUNT BALANCE feature in the set of instances where the target feature FRAUDULENT=*'True'* follows an exponential distribution
  - the distributions of values taken by the ACCOUNT BALANCE feature in the set of instances where the target feature FRAUDULENT=*'False'* is similar to a normal distribution.
- Once we have selected the distributions the next step is to fit the distributions to the data.

- To fit the exponential distribution we simply compute the sample mean, $\bar{x}$, of the ACCOUNT BALANCE feature in the set of instances where FRAUDULENT=*'True'* and set the λ parameter equal to one divided by $\bar{x}$.
- To fit the normal distribution to the set of instances where FRAUDULENT=*'False'* we simply compute the sample mean and sample standard deviation, s, for the ACCOUNT BALANCE feature for this set of instances and set the parameters of the normal distribution to these values.

Gaussian Naive Bayes Classifier assumes that ALL continuous descriptive features follow a normal (Gaussian) distribution and uses each features' mean and std.deviation for fitting this normal distribution.

Table: Partitioning the dataset based on the value of the target feature and fitting the parameters of a statistical distribution to model the ACCOUNT BALANCE feature in each partition.

ID	...	ACCOUNT BALANCE	FRAUD
1		56.75	true
4		749.50	true
6		928.30	true
10	...	405.72	true
12		223.89	true
13		103.23	true
$\overline{\text{AB}}$		411.22	
$\lambda = {}^{1}!/_{\overline{\text{AB}}}$		0.0024	

ID	...	ACCOUNT BALANCE	FRAUD
2		1 800.11	false
3		1 341.03	false
5		1 150.00	false
7		250.90	false
8		806.15	false
9		1 209.02	false
11		550.00	false
14		758.22	false
15		430.79	false
16		675.11	false
17		1 657.20	false
18		1 405.18	false
19		760.51	false
20		985.41	false
$\overline{\text{AB}}$		984.26	
$sd(\text{AB})$		460.94	

Table: The Laplace smoothed (with $k = 3$) probabilities needed by a naive Bayes prediction model calculated from the dataset in Table 5 [23], extended to include the conditional probabilities for the new ACCOUNT BALANCE feature, which are defined in terms of PDFs.

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = none \mid fr)$	$=$	0.2222	$P(CH = none \mid \neg fr)$	$=$	0.1154
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(CH = current \mid fr)$	$=$	0.3333	$P(CH = current \mid \neg fr)$	$=$	0.2692
$P(CH = arrears \mid fr)$	$=$	0.2222	$P(CH = arrears \mid \neg fr)$	$=$	0.3462
$P(GC = none \mid fr)$	$=$	0.5333	$P(GC = none \mid \neg fr)$	$=$	0.6522
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(GC = coapplicant \mid fr)$	$=$	0.2	$P(GC = coapplicant \mid \neg fr)$	$=$	0.2174
$P(ACC = own \mid fr)$	$=$	0.4667	$P(ACC = own \mid \neg fr)$	$=$	0.6087
$P(ACC = rent \mid fr)$	$=$	0.3333	$P(ACC = rent \mid \neg fr)$	$=$	0.2174
$P(ACC = free \mid fr)$	$=$	0.2	$P(ACC = free \mid \neg fr)$	$=$	0.1739
$P(AB = x \mid fr)$			$P(AB = x \mid \neg fr)$		
$\approx$	E	$\left(\begin{array}{c} x, \\ \lambda = 0.0024 \end{array}\right)$	$\approx$	N	$\left(\begin{array}{c} x, \\ \mu = 984.26, \\ \sigma = 460.94 \end{array}\right)$

Table: A query loan application from the fraud detection domain.

Credit History	Guarantor/ CoApplicant	Accomodation	Account Balance	Fraudulent
paid	guarantor	free	759.07	?

Table: The probabilities, from Table 7 [29], needed by the naive Bayes prediction model to make a prediction for the query $\langle \text{CH} = \textit{'paid'}, \text{GC} = \textit{'guarantor'}, \text{ACC} = \textit{'free'}, \text{AB} = 759.07 \rangle$ and the calculation of the scores for each candidate prediction.

$$
\begin{array}{rclrcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(CH = paid | fr) & = & 0.2222 & P(CH = paid | \neg fr) & = & 0.2692 \\
P(GC = guarantor | fr) & = & 0.2667 & P(GC = guarantor | \neg fr) & = & 0.1304 \\
P(ACC = free | fr) & = & 0.2 & P(ACC = free | \neg fr) & = & 0.1739 \\
P(AB = 759.07 | fr) & & & P(AB = 759.07 | \neg fr) & & \\
\approx E\left(\begin{array}{c} 759.07, \\ \lambda = 0.0024 \end{array}\right) & = & 0.00039 & \approx N\left(\begin{array}{c} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{array}\right) & = & 0.00077
\end{array}
$$

$$
\left(\prod_{k=1}^{m} P(\mathbf{q}[k] | fr)\right) \times P(fr) = 0.0000014
$$

$$
\left(\prod_{k=1}^{m} P(\mathbf{q}[k] | \neg fr)\right) \times P(\neg fr) = 0.0000033
$$

For a continuous probability distribution, P(X=c) = 0 for any constant c. So, we use an approximation as below to avoid zero probability issue where epsilon is just a very small number, like 1 dollar in our example:

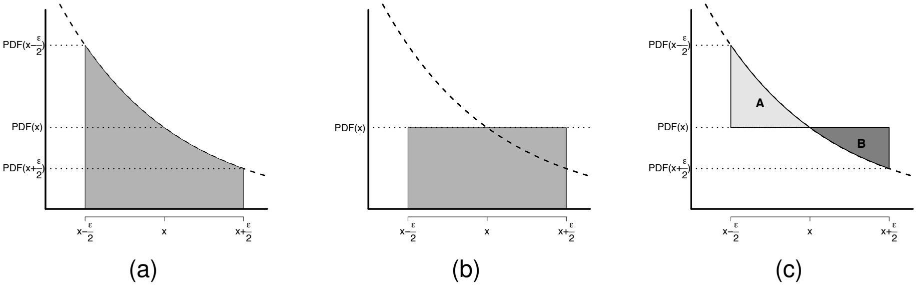

Figure: (a) The area under a density curve between the limits $x - \frac{\epsilon}{2}$ and $x + \frac{\epsilon}{2}$; (b) the approximation of this area computed by $PDF(x) \times \epsilon$; and (c) the error in the approximation is equal to the difference between area A, the area under the curve omitted from the approximation, and area B, the area above the curve erroneously included in the approximation. Both of these areas will get smaller as the width of the interval gets smaller, resulting in a smaller error in the approximation.

Continuous Features: Binning

- In Section 3.6.2 we explained two of the best known binning techniques **equal-width** and **equal-frequency**.
- We can use these techniques to *bin* continuous features into categorical features
- In general we recommend **equal-frequency binning**.

For equal-frequency binning, we can use Panda's "qcut" function.

Table: The dataset from a loan application fraud detection domain with a second continuous descriptive feature added: LOAN AMOUNT

ID	CREDIT HISTORY	GUARANTOR/ COAPPLICANT	ACCOMMODATION	ACCOUNT BALANCE	LOAN AMOUNT	FRAUD
1	current	none	own	56.75	900	true
2	current	none	own	1 800.11	150 000	false
3	current	none	own	1 341.03	48 000	false
4	paid	guarantor	rent	749.50	10 000	true
5	arrears	none	own	1 150.00	32 000	false
6	arrears	none	own	928.30	250 000	true
7	current	none	own	250.90	25 000	false
8	arrears	none	own	806.15	18 500	false
9	current	none	rent	1 209.02	20 000	false
10	none	none	own	405.72	9 500	true
11	current	coapplicant	own	550.00	16 750	false
12	current	none	free	223.89	9 850	true
13	current	none	rent	103.23	95 500	true
14	paid	none	own	758.22	65 000	false
15	arrears	none	own	430.79	500	false
16	current	none	own	675.11	16 000	false
17	arrears	coapplicant	rent	1 657.20	15 450	false
18	arrears	none	free	1 405.18	50 000	false
19	arrears	none	own	760.51	500	false
20	current	none	own	985.41	35 000	false

Table: The Loan Amount continuous feature discretized into 4 equal-frequency bins. Here, we first need to sort from smallest to largest.

ID	Loan Amount	Binned Loan Amount	Fraud
15	500	bin1	false
19	500	bin1	false
1	900	bin1	true
10	9,500	bin1	true
12	9,850	bin1	true
4	10,000	bin2	true
17	15,450	bin2	false
16	16,000	bin2	false
11	16,750	bin2	false
8	18,500	bin2	false

ID	Loan Amount	Binned Loan Amount	Fraud
9	20,000	bin3	false
7	25,000	bin3	false
5	32,000	bin3	false
20	35,000	bin3	false
3	48,000	bin3	false
18	50,000	bin4	false
14	65,000	bin4	false
13	95,500	bin4	true
2	150,000	bin4	false
6	250,000	bin4	true

- Once we have discretized the data we need to record the raw continuous feature threshold between the bins so that we can use these for query feature values.

Table: The thresholds used to discretize the LOAN AMOUNT feature in queries.

Bin Thresholds		
	Bin1	$\leq 9,925$
$9,925 <$	Bin2	$\leq 19,250$
$19,225 <$	Bin3	$\leq 49,000$
$49,000 <$	Bin4	

Table: The Laplace smoothed (with $k = 3$) probabilities needed by a naive Bayes prediction model calculated from the fraud detection dataset. Notation key: FR = FRAUD, CH = CREDIT HISTORY, AB = ACCOUNT BALANCE, GC = GUARANTOR/COAPPLICANT, ACC = ACCOMMODATION, BLA = BINNED LOAN AMOUNT.

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = none \mid fr)$	$=$	0.2222	$P(CH = none \mid \neg fr)$	$=$	0.1154
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(CH = current \mid fr)$	$=$	0.3333	$P(CH = current \mid \neg fr)$	$=$	0.2692
$P(CH = arrears \mid fr)$	$=$	0.2222	$P(CH = arrears \mid \neg fr)$	$=$	0.3462
$P(GC = none \mid fr)$	$=$	0.5333	$P(GC = none \mid \neg fr)$	$=$	0.6522
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(GC = coapplicant \mid fr)$	$=$	0.2	$P(GC = coapplicant \mid \neg fr)$	$=$	0.2174
$P(ACC = own \mid fr)$	$=$	0.4667	$P(ACC = own \mid \neg fr)$	$=$	0.6087
$P(ACC = rent \mid fr)$	$=$	0.3333	$P(ACC = rent \mid \neg fr)$	$=$	0.2174
$P(ACC = free \mid fr)$	$=$	0.2	$P(ACC = free \mid \neg fr)$	$=$	0.1739
$P(AB = x \mid fr)$			$P(AB = x \mid \neg fr)$		
$\approx E\begin{pmatrix} x, \\ \lambda = 0.0024 \end{pmatrix}$			$\approx N\begin{pmatrix} x, \\ \mu = 984.26, \\ \sigma = 460.94 \end{pmatrix}$		
$P(BLA = bin1 \mid fr)$	$=$	0.3333	$P(BLA = bin1 \mid \neg fr)$	$=$	0.1923
$P(BLA = bin2 \mid fr)$	$=$	0.2222	$P(BLA = bin2 \mid \neg fr)$	$=$	0.2692
$P(BLA = bin3 \mid fr)$	$=$	0.1667	$P(BLA = bin3 \mid \neg fr)$	$=$	0.3077
$P(BLA = bin4 \mid fr)$	$=$	0.2778	$P(BLA = bin4 \mid \neg fr)$	$=$	0.2308

Table: A query loan application from the fraud detection domain.

Credit History	Guarantor/ CoApplicant	Accomodation	Account Balance	Loan Amount	Fraudulent
paid	guarantor	free	759.07	8,000	?

Table: The relevant smoothed probabilities, from Table 13 [37], needed by the naive Bayes model to make a prediction for the query $\langle \text{CH} = \textit{'paid'}, \text{GC} = \textit{'guarantor'}, \text{ACC} = \textit{'free'}, \text{AB} = 759.07, \text{LA} = 8\,000 \rangle$ and the calculation of the scores for each candidate prediction.

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(ACC = free \mid fr)$	$=$	0.2	$P(ACC = free \mid \neg fr)$	$=$	0.1739
$P(AB = 759.07 \mid fr)$			$P(AB = 759.07 \mid \neg fr)$		
$\approx E\left(\begin{array}{c} 759.07, \\ \lambda = 0.0024 \end{array}\right)$	$=$	0.00039	$\approx N\left(\begin{array}{c} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{array}\right)$	$=$	0.00077
$P(BLA = bin1 \mid fr)$	$=$	0.3333	$P(BLA = bin1 \mid \neg fr)$	$=$	0.1923

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.000000462$$

$$\left(\prod_{k=1}^{n} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.000000633$$

Summary

- Naive Bayes models can suffer from zero probabilities of relatively rare events. **Smoothing** is an easy way to combat this.
- Two ways to handle continuous features in probability-based models are: **Probability density functions** and **Binning**
- Using probability density functions requires that we match the observed data to an existing distribution.
- Although binning results in information loss it is a simple and effective way to handle continuous features in probability-based models.
- ~~Bayesian network representation is generally more compact than a full joint distribution, yet is not forced to assert global conditional independence between all descriptive features.~~