Encoding Numerical Columns

- Discretization / Binning
  - Unsupervised
    - Equal Width
    - Equal Frequency
    - K - Means Binning
  - Supervised
    - Decision Tree Binning
  - Custom
- Binarization

Discretization

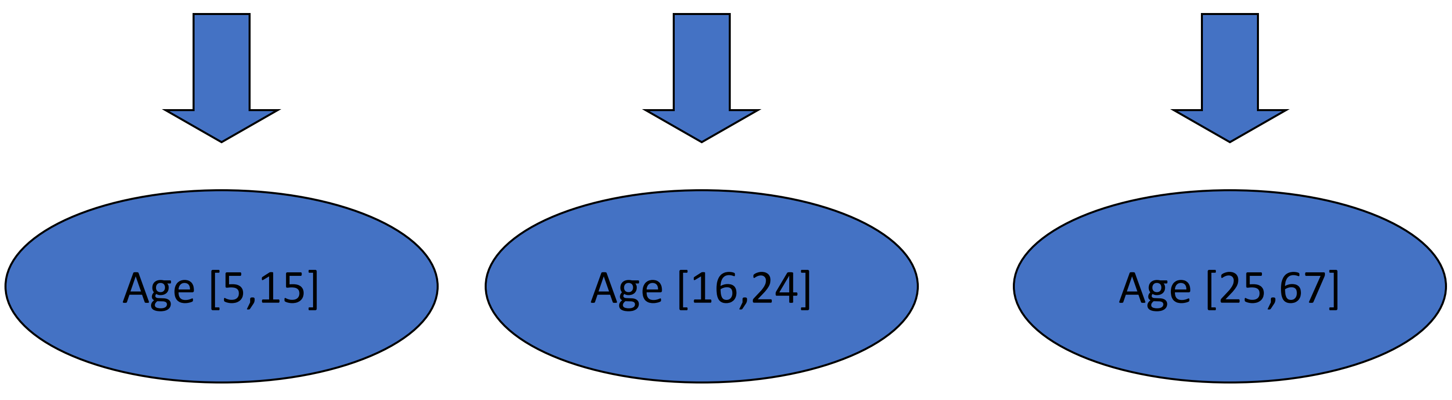
- Discretization:
  - divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.

- Experimental results indicate that with discretization
  - data size can be reduced
  - classification accuracy can be improved

Discretization

- Store only the interval labels
- Important for association rules and classification

age	5	6	6	9	…	15	16	16	17	20	…	24	25	41	50	65	…	67
own a car	0	0	0	0	…	0	1	0	1	1	…	0	1	1	1	1	…	1



Entropy-Based Discretization (Supervised)

- Given a set of samples S, if S is partitioned into two intervals S1 and S2 using boundary T, the entropy after partitioning is

$$E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization.
- The process is recursively applied to partitions obtained until some stopping criterion is met, e.g.,

$$Ent(S) - E(T,S) > \delta$$

Example 1

ID	1	2	3	4	5	6	7	8	9
Age	21	22	24	25	27	27	27	35	41
Grade	F	F	P	F	P	P	P	P	P

- Let Grade be the class attribute. Use entropy-based discretization to divide the range of ages into differen

(22+24) / 2 = 23

- There are 6 p ble boundaries. They are 21.5, 23, 24.5, 26, 31, and 38

(21+22) / 2 = 21.5

- Let us consider the boundary at T = 21.5.

Let S1 = {21}

S2 = {22, 24, 25, 27, 27, 27, 35, 41}

Example 1 (cont')

ID	1	2	3	4	5	6	7	8	9
Age	**21**	**22**	**24**	**25**	**27**	**27**	**27**	**35**	**41**
Grade	**F**	**F**	**P**	**F**	**P**	**P**	**P**	**P**	**P**

- The number of elements in *S*1 and *S*2 are:

  |*S*1| = 1

  |*S*2| = 8

- The entropy of *S*1 is

$$Ent(S_1) = -P(Grade = \mathrm{F}) \times \log_2 P(Grade = \mathrm{F}) - P(Grade = \mathrm{P}) \times \log_2 P(Grade = \mathrm{P})$$
$$= -(1) \times \log_2 (1) - (0) \times \log_2 (0)$$
$$=$$

- The entropy of *S*2 is

$$Ent(S_2) = -P(Grade = \mathrm{F}) \times \log_2 P(Grade = \mathrm{F}) - P(Grade = \mathrm{P}) \times \log_2 P(Grade = \mathrm{P})$$
$$= -(2) \times \log_2 (2) - (6) \times \log_2 (6)$$
$$=$$

Example 1 (cont')

- Hence, the entropy after partitioning at *T* = 21.5 is

$$E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

$$= \frac{|1|}{|9|} Ent(S_1) + \frac{|8|}{|9|} Ent(S_2)$$

$$= \ldots$$

Example 1 (cont')

- The entropies after partitioning for all the boundaries are:

T = 21.5 = E(S,21.5)

T = 23 = E(S,23)

.

.

T = 38 = E(S,38)

Select the boundary with the smallest entropy

Suppose best is T = 23

Now recursively apply entropy discretization onto both partitions

ID	1	2	3	4	5	6	7	8	9
Age	21	22	24	25	27	27	27	35	41
Grade	F	F	P	F	P	P	P	P	P

Example 2

Age	Buy
10	No
15	No
18	Yes
19	Yes
24	No
29	Yes
30	Yes
31	Yes
40	No
44	No
55	No
64	No

Split point = 35.5

Recursively find the best partition that minimizes entropy

	Yes	No
< 35.5	5	3
>= 35.5	0	4

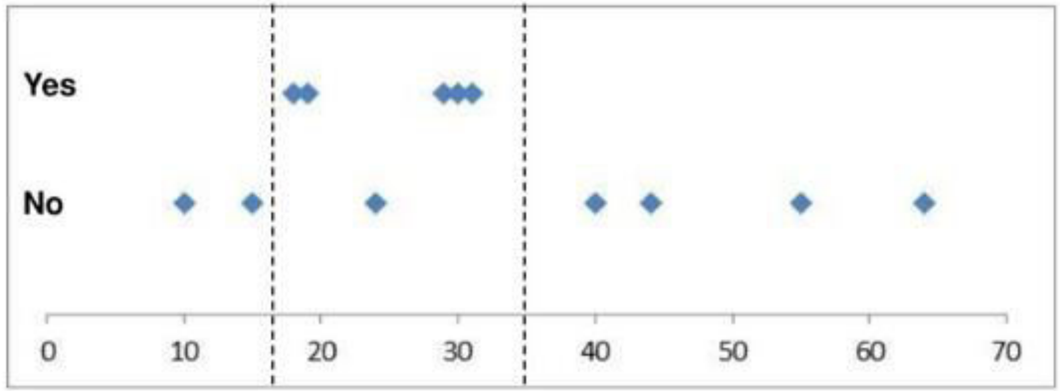
$$E_1 = -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} = 0.9544$$

$$E_2 = -\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4} = 0$$

$$E_T = \frac{8}{12}E_1 + \frac{4}{12}E_2 = 0.6363$$

Supervised Discretization

Age	Buy
10	No
15	No
18	Yes
19	Yes
24	No
29	Yes
30	Yes
31	Yes
40	No
44	No
55	No
64	No



Supervised discretization:

	Yes	No
< 16.5	0	2
(16.5,35.5)	5	1
> 35.5	0	4

In supervised discretization, our goal is to ensure that each bin contains data points from one class.