Lecture 18: Entropy, Information Gain, Decision Trees

Recap

- Evaluation metrics

Information, Entropy, Mutual Information

Information

- The Sun will rise in the east tomorrow and next week
- My phone will ring in the next one hour
- It will snow in Manipal this winter

- Which of these statements contains the maximum information and least information?

What is information?

- Electronic communication between two places
- 1 bit can send 2 states – 0 and 1: 2^1
- 2 bit can send 4 states – 00, 01, 10, 11 : 2^2
- 3 bits can send 2^3 states
- N bits can send 2^N states

- Message space of N bits is 2^N

- Let message space be x
- Number of bits for message space $log_2 x$
- Number of bits required to transmit a message is called information
- Aka Shannon Information Content

$$I(x) = log_2 x$$

- Unit of Shannon Information content is bits

Information contained in English alphabets

- Message space x = 26

$$I(English \ \ alphabets) = log_2 26 = 4.7$$

- 4.7 bits needed to communicate English alphabet
- Only if all English alphabets were equi-probable
- This is not the case

Information as a measure of uncertainty

- Number of bits $I(x) = log_2 x$
- If all events are equi-probable p = 1/x implies x = 1/p

$$I(x) = log_2 \frac{1}{p} = log_2 1 - log_2 p = 0 - log_2 p = -log_2 p$$

- $log_2 x, log x$ are monotonic.
- Replace with natural logarithm $I(x) = -log_e p(x)$
- Unit is nats (instead of bits)

- Revisit information contained in sun rising in east

Entropy = Average Information

$$I(x) = -log_e p(x)$$

$$\mathbb{E}[I(x)] = -\mathbb{E}[log p(x)] = -\sum p(x) log p(x)$$

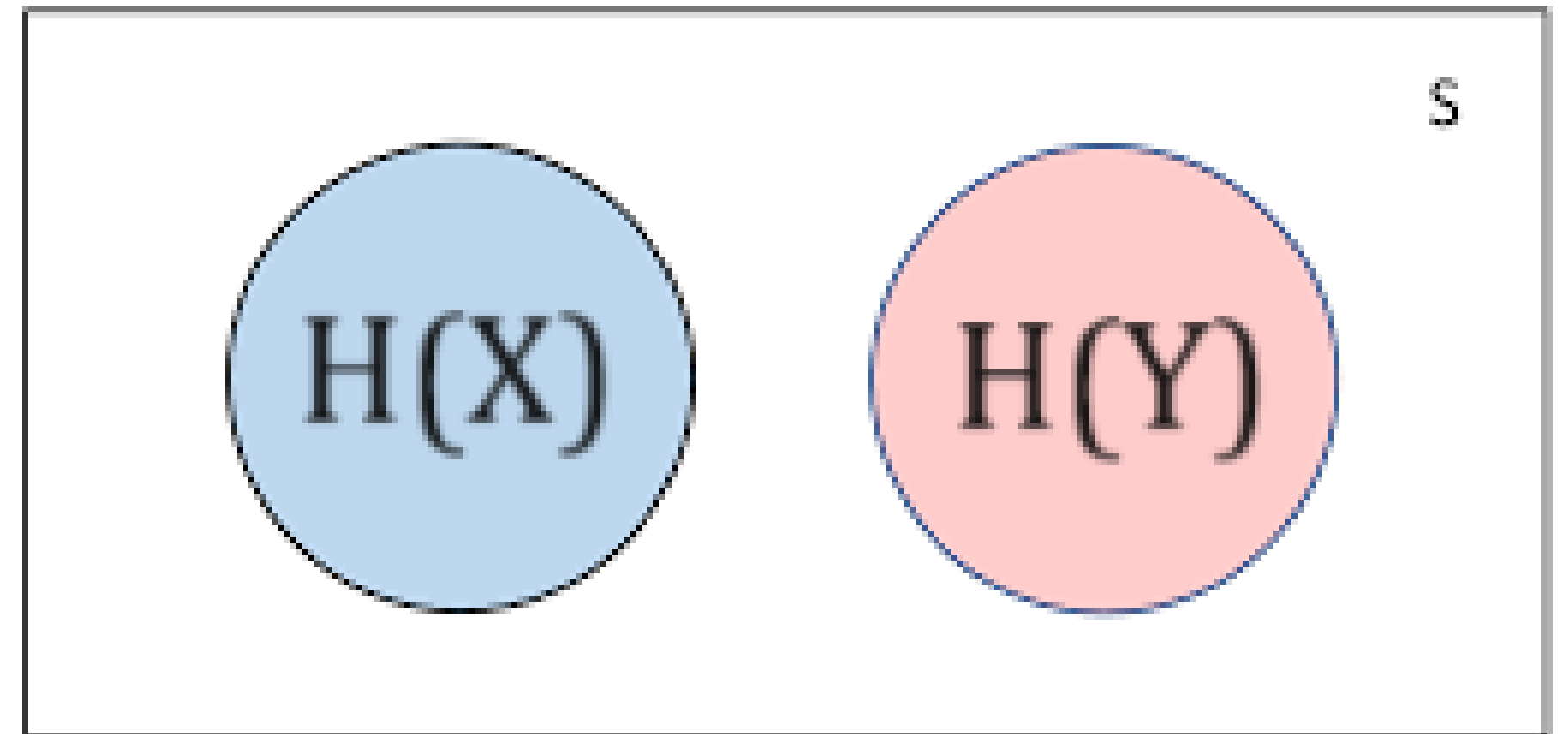$$H = \mathbb{E}[I(x)] = -\sum p(x) log p(x)$$

- From communication perspective:
  - Number of bits needed to communicate

- From ML perspective:
  - Amount of info contained in features & target

Warning: Terminology overload

- Joint Entropy
- Conditional Entropy
- Mutual Information
- Information Gain

- Cross Entropy
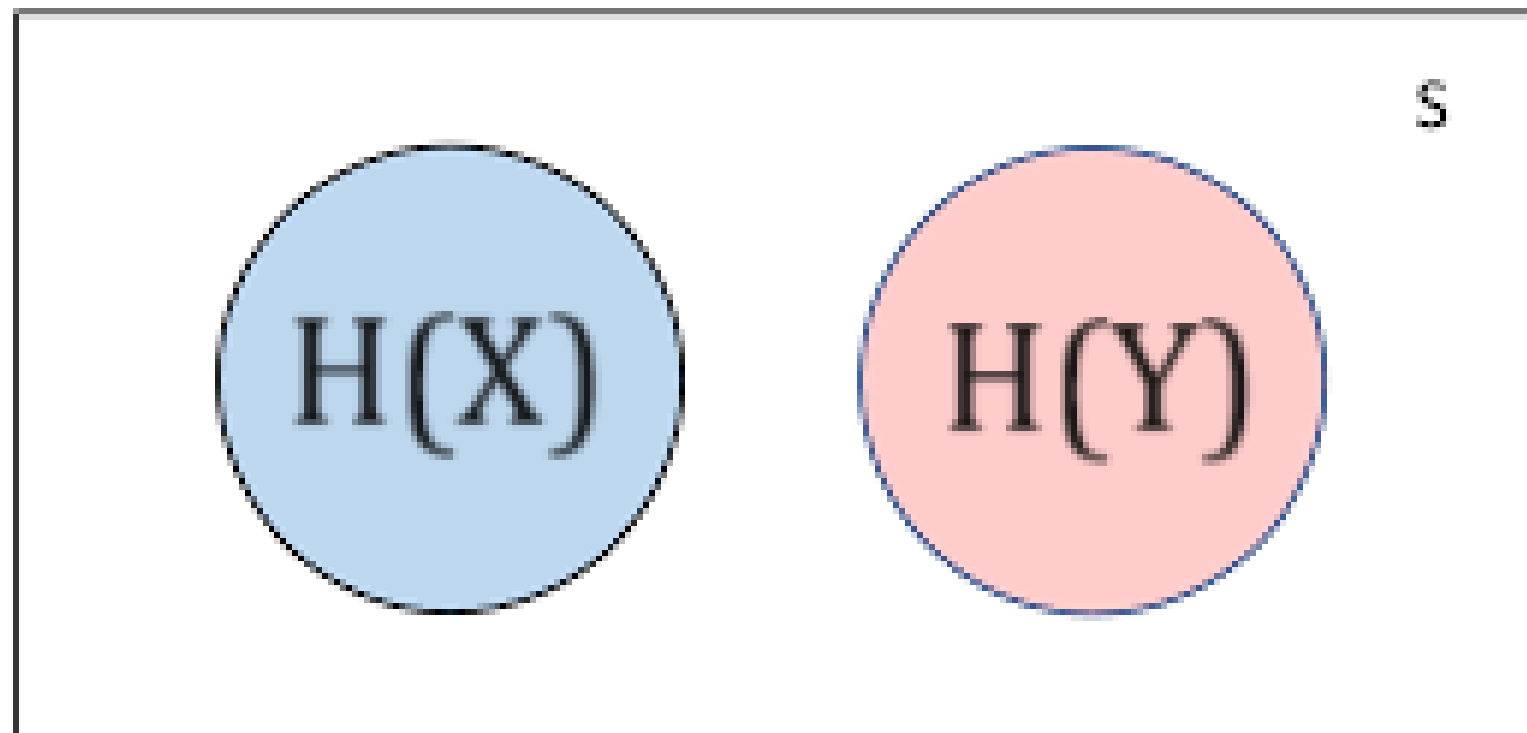- KL Divergence

Independent events and Joint Entropy

- $P(X, Y) = P(X)\ P(Y)$
- $\log P(X,Y) = \log P(X) + \log P(Y)$
- $H(X,Y) = H(X) + H(Y)$
- Independent probabilities are multiplicative
- Independent Information is Additive
  - Information adds up.
  - That's logical
- Joint entropy is
  - Not intersection

Joint Entropy Proof

$$H(X) =$$

$$-\sum p(X = x)log(p(X = x))$$

$$= -\sum p(x)logp(x)$$

$$H(X,Y) = -\sum_{x,y} p(x,y)logp(x,y)$$

$$= -\sum_{x,y} p(x)p(y)log(p(x)p(y))$$

$$= -\sum_{x,y} p(x)p(y)(logp(x) + logp(y))$$

$$= -\sum_{x} p(x)logp(x) - \sum_{y} p(y)logp(y)$$

$$= H(X) + H(Y)$$

H(X) H(Y)

s

Dependent events and Joint Entropy

- $P(X, Y) = P(X)\, P(Y|X)$
- $\log P(X,Y) = \log P(X) + \log P(Y|X)$
- $H(X,Y) = H(X) + H(Y|X)$
- $= H(Y) + H(X|Y)$
- X1, X2.. Features
- Y is target variable
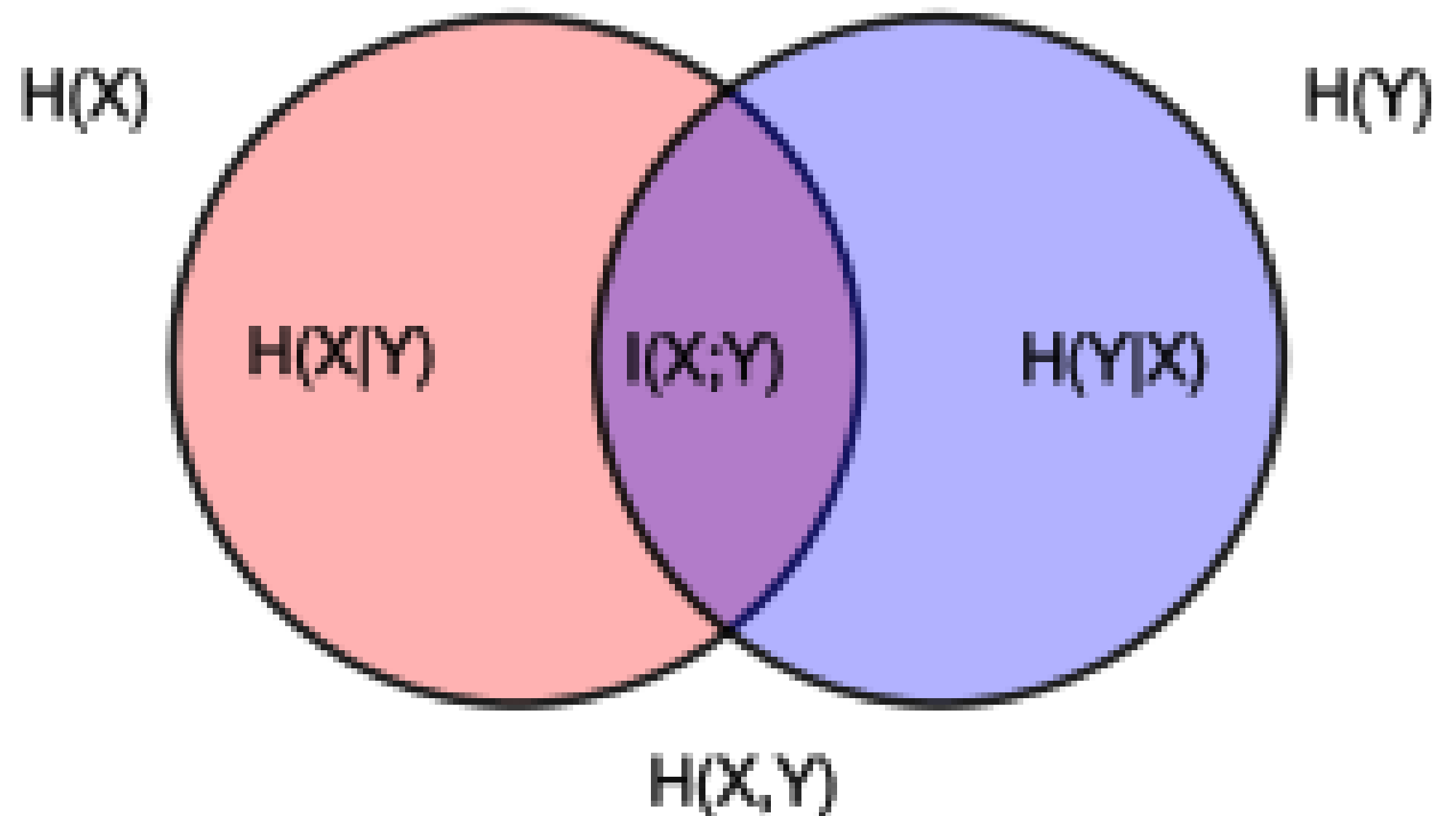
Mutual Information (MI)

- MI is the common info between two features in ML
- MI represented as I(X,Y) - between feature & target
- I(X,Y) = H(X) + H(Y) − H(X,Y)
- = H(Y) − H(Y|X)
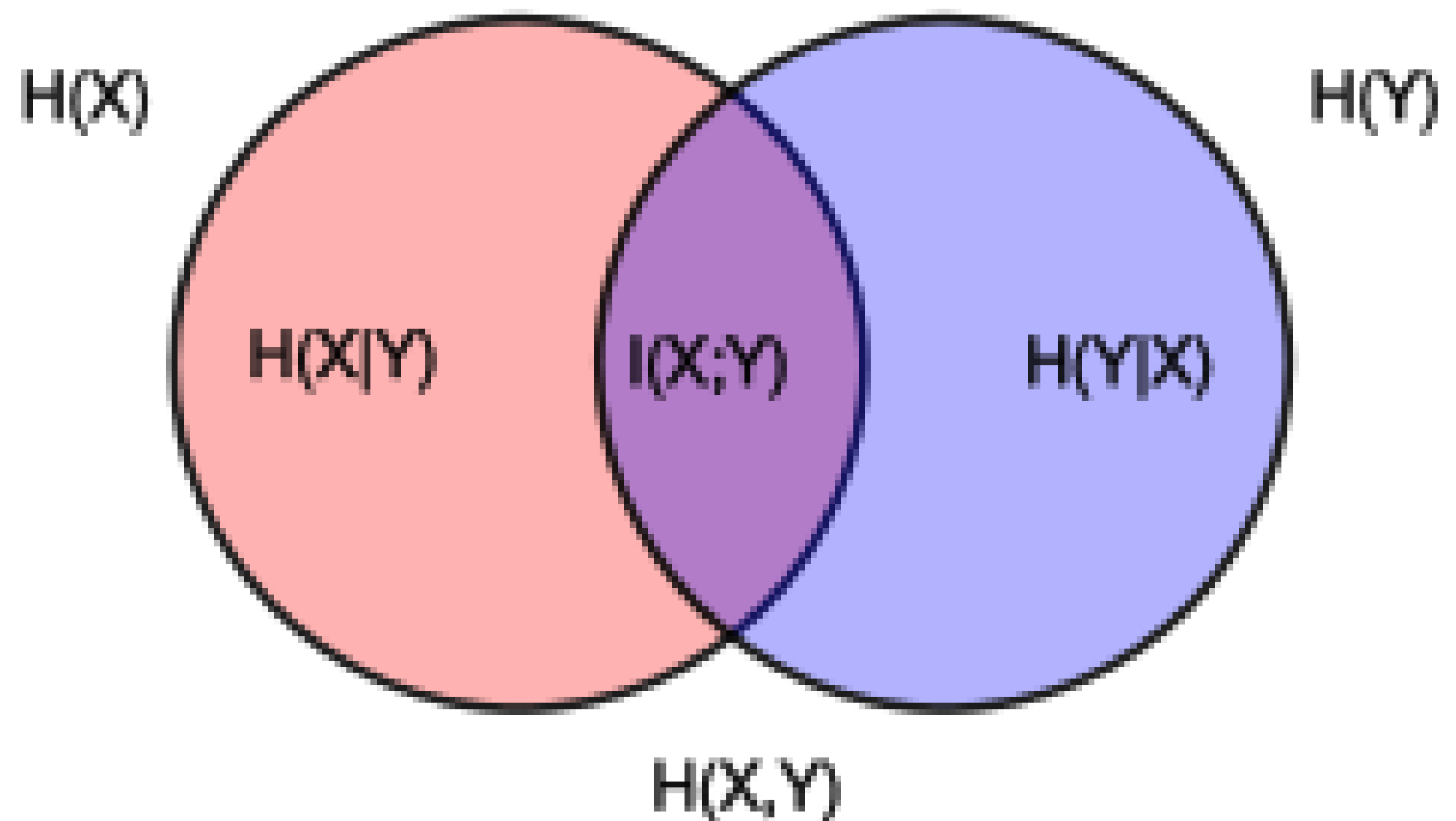
Using Entropy and Mutual Information in Decision Trees

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

- Calculate Entropy of dataset target variable

$$H(S) = -\frac{5}{14}log(\frac{5}{14}) - \frac{9}{14}log(\frac{9}{14})$$

$$-(0.35)(-2.63) - (0.64)(-0.44)$$

$$= 0.92 + 0.28 = 1.2$$

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

$$H(S) = 1.2$$

$$H(S|Humidity = High) =$$
$$-\frac{4}{7}log(\frac{4}{7}) - \frac{3}{7}log(\frac{3}{7}) = 0.66$$

$$H(S|Humidity = Normal) =$$
$$-\frac{6}{7}log(\frac{6}{7}) - \frac{1}{7}log(\frac{1}{7}) = 0.4$$

$$H(S|Humidity) = \frac{7}{14}0.66 + \frac{7}{14}0.4$$
$$= 0.53$$

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

$$H(S) = 1.2$$

$$H(S|Humidity) = 0.53$$

$$Information\ Gain$$
$$IG(Humidity) = I(Humidity, Y)$$
$$= H(S) - H(S|Humidity)$$
$$= 1.2 - 0.53 = 0.63$$

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

- Calculate
  - IG(Outlook)
  - IG(Humidity)
  - IG(Wind)
- Split on feature subject to max of the three

19

Using Entropy concepts

- Decision Tree
- Feature Selection (Exercise)
- sklearn.feature_selection
  - SelectPercentile
  - SelectKBest

Terminology not covered in this lecture

- Cross Entropy
- KL Divergence
- Important for softmax
- Important for deep learning & neural networks

Further reading

- https://colah.github.io/posts/2015-09-Visual-Information/
- https://charlesfrye.github.io/stats/2016/03/29/info-theory-surprise-entropy.html