Ethan Hadley

1: During decision induction, an essential step, which is repeated at every none-left node from root down, is to select the attribute (among attributes that are still available at a particular branch), that will provide the most gain of information (also known as entropy). Discuss the bias of this approach and what C4.5, a successor of ID3, did as an attempt to overcome that bias. (15)

The maximum entropy distribution is the uniform distribution. If an attribute has many possible values, each of which occurrs rarely (or only once in the extreme case), its entropy will be very high, and ID3 will priotitize it for selection. This could be undesirable if the attribute, while having high entropy in and of itself, is uncorrelated or only weakly correlated with the attribute we care about predicting. C4.5 instead prioritizes based on entropy after adjusting for the evenness of the attribute's split, called the Gain Ratio. If the attribute splits the data into many evenly sized groups, the GR will be lower. If the groups are uneven or there are fewer of them, the GR will be higher. This allows us to make more compact and accurate trees which properly priotize attributes that are more likely to generalize rather than memorize.

2:

- 1: Supervised, Class
- 2: Gini
- 3: overfitting, leaf
- 4: focus?
- 5: Supervised? Missing value?
- 6: clustering, unsupervised
- 7: "_______ is a commonly used method for smooth data by fitting them into a mathematical function." is not a sentence. idk if this means it's a method for smoothing data, or a method that uses smooth data. Assuming it is the latter, my guess is 'regression'.
- 8: discretization

3:

P(play) = 7 / 14 = 0.5 P(sunny | play) = 1 / 7 = 0.143 P(sunny | !play) = 4 / 7 = 0.571 P(windy | play) = 2 / 7 = 0.286 P(windy | !play) = 4 / 7 = 0.571 now calculating the mean and variance of the humidity variable given that we play:
$$\large \mu = \frac{65 + 50 + 70 + 50 + 75 + 70 + 50}{7} = 61.42$$ $$\large \sigma^2 = \frac{(65-61.42)^2+(50-61.42)^2+(70-61.42)^2+(50-61.42)^2+(75-61.42)^2+(70-61.42)^2+(50-61.42)^2}{7} = 105.10$$ and given that we dont play: $$\large \mu = \frac{95 + 90 + 85 + 90 + 70 + 40 + 65}{7} = 76.43$$ $$\large \sigma^2 = \frac{(95-76.42)^2+(90-76.42)^2+(85-76.42)^2+(90-76.42)^2+(70-76.42)^2+(40-76.42)^2+(65-76.42)^2}{7} = 326.53$$

so P(humidity = 70 | play) = $$\large P(h) = \frac{1}{\sqrt{2\pi\sigma^2}}exp(-\frac{(70 - \mu)^2}{2\sigma^2})$$ $$= \frac{1}{\sqrt{2\pi105.10}}exp(-\frac{(70 - 61.42)^2}{2*105.10}) = 0.0274$$ and P(humidity = 70 | !play) = $$\frac{1}{\sqrt{2\pi326.53}}exp(-\frac{(70 - 76.42)^2}{2*326.53}) = 0.0207$$

So P(sunny & windy & humidity=70 | play) = 0.143 * 0.286 * 0.0274 = 0.00112
and P(sunny & windy & humidity=70 | !play) = 0.571 * 0.571 * 0.0207 = 0.0067

And since the prior probability of play is 0.5, it makes no difference and so we can say that the probability of not playing is greater, so we predict "don't play".

4:

"if humidity == low then play == yes else play == false" has 80% coverage, being correct for 4 out of the 5 occurrences of humidity == low.

5: lol

- 1: Sundar (not sundai) Pichai, Deep Mind
- 2: Satya Nadella, OpenAI, ChatGPT, Sam Altman
- 3: NVIDIA, Jensen Huang, Lisa Su, AMD
- 4: Boring