David Varodayan
CS109

Section #4

1. **Are we due for an earthquake?** After the class in which we talked about the probability of earthquakes, a student asked: "Doesn't the probability of an earthquake happening change based on the fact that we haven't had one for a while?" Let's explore! Recall the USGS rate of earthquakes of magnitude 8+ in California is $\lambda = 0.002$ earthquakes per year.

   a. What is the probability of no 8+ earthquakes in four years after the 1906 earthquake (recall that earthquakes are exponentially distributed)?

   b. What is the probability of no 8+ earthquakes in the 117 years between 1906 and four years from now?

   c. What is the probability of no 8+ earthquakes in the 117 years between 1906 and four years from now *given* that there have been no earthquakes in the last 113 years?

   d. Did you notice anything interesting? Would this work for any value of λ?

2. **Fairness in AI.** In their 2018 paper "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," Joy Buolamwini and Timnit Gebru showed that commercial gender classifiers performed significantly worse on face images of darker-skinned females (error rates up to 34.7%) than lighter-skinned males (maximum error rate of 0.8%.) This disparity may result from training the classifiers on unbalanced datasets. To evaluate the classifiers, the authors designed their own dataset by collecting photos of national parliamentarians from three African countries and three European ones.

   The probability table below shows the joint distribution of the dataset between two random variables: the demographic (D) of the photo subject and their country (C).

Demographic	South Africa	Senegal	Rwanda	Sweden	Finland	Iceland
Darker Female	0.12	0.05	0.04	0.01	0	0
Darker Male	0.15	0.07	0.02	0.01	0	0
Lighter Female	0.02	0	0	0.12	0.06	0.02
Lighter Male	0.05	0	0	0.14	0.09	0.03

   a. What is the marginal probability distribution for demographic D? Provide your result as a mapping from values that D can take to probabilities.

   b. What is the conditional probability of country given that the subject is a lighter female, $P(C|D = \text{Lighter Female})$? Provide your result as a mapping from values that C can take to probabilities. Is this mapping a probability distribution?

c. What is the conditional probability that the subject is from Senegal given their demographic, $P(C = \text{Senegal}|D)$? Provide your answer as a mapping from values that D can take to probabilities. Is this mapping a probability distribution?

d. What are the pitfalls in using this dataset for a purpose beyond what the authors intended?

3. **ReCaptcha.** Based on browser history, Google believes that there is a 0.2 probability that a particular visitor to a website is a robot. They decide to give the visitor a reCaptcha:
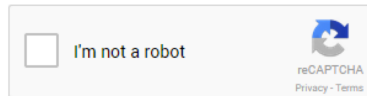
Google presents the visitor with a 10 mm by 10 mm box. The visitor must click inside the box to show that they are not a robot. You have observed that robots click uniformly in the box. However, the distance location of a human click has X location (mm from the left) and the Y location (mm from the top) distributed as independent normals both with mean $\mu = 5$ and variance $\sigma^2 = 4$.

a. What is the probability density function of a robot clicking $X = x$ mm from the left of the box and $Y = y$ mm from the top of the box?

b. What is the probability density function of a human clicking $X = x$ mm from the left of the box and $Y = y$ mm from the top of the box?

c. The visitor clicks in the box at ($x = 6$ mm, $y = 6$ mm). What is Google's new belief that the visitor is a robot?