Learning Machine Learning: Statistics & Probability 1/30/20

What does learning mean?

→ considering information & make inferences based off of that information

↳ using logic & probability

deductive vs. inferential (inductive) reasoning

· deductive: absolute logic

ex) $A \to B \Rightarrow \bar{B} \to \bar{A}$ (converse) ($\bar{A}$ = not A)

· inductive reasoning: $B \not\Rightarrow A$, $\bar{A} \not\Rightarrow B$, absolutely

theory of logic: quantify all learning that isn't deductive

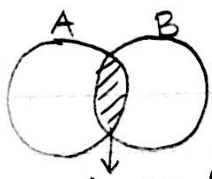
· inductive reasoning is mostly everyday reasoning

$A \to B \Rightarrow P(B|A) = 1$: if A is true then B is true

sum rule: $P(A+B) = P(A) + P(B) - P(AB)$

· $P(A+B)$ is $P(A$ or $B)$

· $P(AB)$ is $P(A$ and $B)$




A and B (remove double counting)

product rule: $P(AB) = P(A|B)P(B) = P(B|A)P(A)$

→ these rules allow for the creation of a self-consistent theory of probability

↳ two processes will come to the same result

⇒ sum + product rules give Bayes theorem

$P(D|X) = P(D)P(X|D)/P(X)$

· D is event & X is evidence

· $P(D)$: prior – encodes previous knowledge

· $P(X)$: evidence (marginal likelihood), normalization

· $P(X|D)$: likelihood – how consistent is data w/ observation

· $P(D|X)$: posterior

ex) $P(F|N) = P(F)P(N|F)/P(N)$

- F: event happening (sky is falling)
- N: evidence → newscaster telling the truth
  - ↳ $P(N) = P(F)P(N|F) + P(\bar{F})P(N|\bar{F})$
- $P(F) = 10^{-9}$ (prior)
- $P(N|F) = 1$
- $P(F)P(N|F) = 10^{-9}(1)$
- $P(\bar{F})P(N|\bar{F}) = 0$

$$\left.\right\} \frac{10^{-9} \cdot 1}{10^{-9} \cdot 1 + 1 \cdot 0} = 1$$

↳ trustworthy newscaster

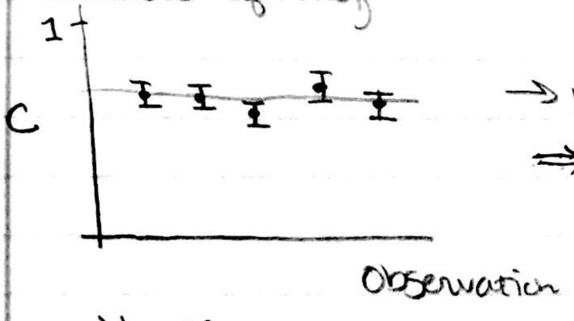
- $P(\bar{F})P(N|\bar{F}) = 0.1$ → newscaster lies 10% of the time
  - ↳ $\frac{10^{-9} \cdot 1}{10^{-9} \cdot 1 + 1(0.1)} = 10^{-8}$
- $P(N|F) = 0.1$ → newscaster tells truth 10% of the time
  - ↳ $\frac{10^{-9}(0.1)}{10^{-9}(0.1) + 1(0.9)} = \frac{10^{-10}}{.9} \gtrsim 10^{-10}$

→ as $P(N|F) \to 0 \Rightarrow P(F|N) \to 0$ faster

- Bayes' theorem tells us how to update our information based on same intake of information

Models & parameters

ex) Cuteness of dog




→ what is the true "cuteness"?

⇒ fit line to observation → need model

$Y_i = C + \varepsilon$

- data ($Y_i$) is equal to true cuteness ($C$) w/ error ($\varepsilon$)
  - $\varepsilon$ is normally distributed

Scanned with CamScanner

1/30/20

ex (cont) likelihood of one point given to true value $c$ is related to the data point's distance to the true value

$P(y_i|c) = e^{-(y_i-c)^2/2\sigma^2} \,/\, \sqrt{2\pi\sigma^2}$ for one data point

$\Rightarrow \mathcal{L}(c) = P(\vec{y}|c) = \prod_i^N 1/\sqrt{2\pi\sigma}\; e^{-(y_i-c)^2/2\sigma^2}$ for likelihood over all data pts.

· Bayes thm tells us : $P(c|\vec{y}) = P(\vec{y})\,P(\vec{y}|c) \,/ \int P(c')\,P(\vec{y}|c')\,dc'$

$\rightarrow$ for no prior $(P(\vec{y}) = 1) \Rightarrow P(c|\vec{y}) = P(\vec{y}|c)$ (frequentist approach)

$= (2\pi\sigma^2)^{-N/2}\, e^{-1/2\sigma^2(\sum_i (y_i-c)^2)} = (2\pi\sigma^2)^{-N/2}\, e^{-1/\sigma^2(\sum_i y_i^2 - 2\sum_i y_i c + Nc^2)}$

$\rightarrow$ probability distribution of $c$ given $\vec{y}$

$= 1/\sqrt{2\pi\sigma^2/N}\; e^{-N(\bar{y}-c)^2/2\sigma^2}$ (gaussian distribution of $c$)