2) "I have 3 kids. Each of them rolled a pair of dice. I have at least one boy who rolled (1,1)."

What is the prob. that I have 3 boys?

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A)\,\mathcal{P}(A)}{\mathcal{P}(B)}$$

BBB = all 3 kids are boys

$\tilde{B}_s$ = at least one is a boy who rolled (1,1)

$$\mathcal{P}(BBB|\tilde{B}_s) = \frac{\mathcal{P}(\tilde{B}_s|BBB)\,\mathcal{P}(BBB)}{\mathcal{P}(\tilde{B}_s)}$$

$$\mathcal{P}(BBB) = \frac{1}{8} \qquad \mathcal{P}(\tilde{B}_s|BBB) = 1 - \overbrace{(1-\epsilon)^3}^{\text{prob. no one rolled } (1,1)}$$

ϵ = prob. of $(1,1) = \frac{1}{36}$

$1-\epsilon$ = prob. did not roll (1,1)

trick to calc. denominator in Bayes' rule:

$$1 = \sum_A \mathcal{P}(A|B) = \sum_A \frac{\mathcal{P}(B|A)\,\mathcal{P}(A)}{\mathcal{P}(B)}$$

$$1 = \frac{1}{\mathcal{P}(B)} \sum_A P(B|A)\,P(A)$$

$$\Rightarrow \mathcal{P}(B) = \sum_A \mathcal{P}(B|A)\,\mathcal{P}(A)$$

$$P(\tilde{B}_s) = \sum_{\substack{XYZ \\ \in \text{ all 3} \\ \text{kids combos}}} P(\tilde{B}_s \mid XYZ)\, \underbrace{P(XYZ)}_{1/8}$$

$$P(\tilde{B}_s \mid GGG) = 0$$

$$P(\tilde{B}_s \mid BGG) = \epsilon$$

$$\text{"} \quad GBG = \epsilon$$

$$GGB$$

$$P(\tilde{B}_s \mid BBG) = 1-(1-\epsilon)^2$$

$$\vdots$$

algebra : $P(\tilde{B}_s) = \frac{1}{8}\epsilon(12 - 6\epsilon + \epsilon^2)$

$$\Rightarrow \quad P(BBB \mid \tilde{B}_s) = \frac{1-(1-\epsilon)^3}{\epsilon(12-6\epsilon+\epsilon^2)}$$

$$\epsilon = \frac{1}{36} \Rightarrow \quad \approx 0.247 > \frac{1}{7}$$

limit of small ϵ Taylor expand $\approx \frac{1}{4} - \frac{\epsilon}{8}$

more abstract interpretation

$$\mathcal{P}(BBB \mid \underbrace{\tilde{B}_s}_{\text{data}}) = \underbrace{\left[\frac{\mathcal{P}(\tilde{B}_s \mid BBB)}{\mathcal{P}(\tilde{B}_s)}\right]}_{\text{rule for updating our knowledge}} \underbrace{\mathcal{P}(BBB)}_{\text{"prior": knowledge before knowing } \tilde{B}_s \text{ (hypothesis)}}$$

"posterior": knowledge after accounting for the data $\tilde{B}_s$

$\mathcal{P}(\tilde{B}_s \mid BBB) \Rightarrow$ "likelihood": prob. of data given prior knowledge

$\mathcal{P}(\tilde{B}_s) \Rightarrow$ normalization const.

---

model fitting: dataset $\mathcal{D}$

model w/ some parameters $\vec{w} = (w_1, w_2, \ldots)$

$$\underbrace{\mathcal{P}(\vec{w} \mid \mathcal{D})}_{\text{posterior}} = \frac{\mathcal{P}(\mathcal{D} \mid \vec{w})}{\mathcal{P}(\mathcal{D})} \underbrace{\mathcal{P}(\vec{w})}_{\text{prior}}$$

prior (reflects physical ranges of possible params)

note: to do MAP you only need to max. $\mathcal{P}(\mathcal{D} \mid \vec{w})\, \mathcal{P}(\vec{w})$ + can ignore $\mathcal{P}(\mathcal{D})$

alternatively: $\mathcal{P}(\vec{w}) = \text{const.}$

$\Rightarrow$ no clue about what the params are

Two major applications:

1) find "best" model parameters:

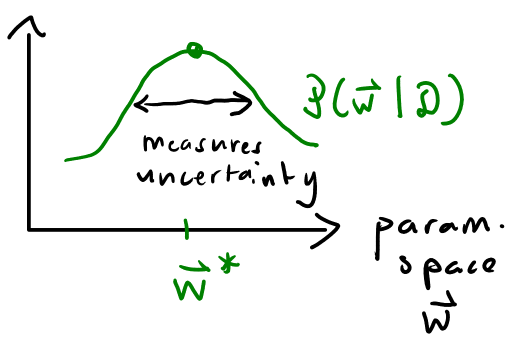
maximize $P(\vec{w} \mid D)$ w/ respect to $\vec{w}$

$\Rightarrow$ $\vec{w}^*$ is the "best" param. set.

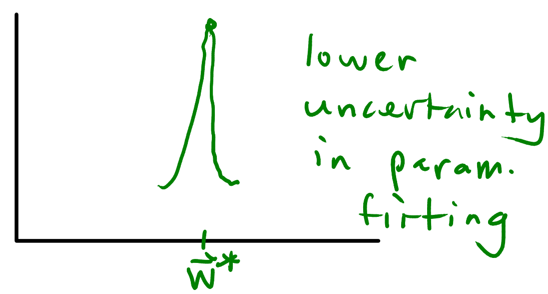
$\Rightarrow$ MAP: maximum a posteriori fitting

(typical model fitting, i.e. training neural networks)

2) find $P(\vec{w} \mid D)$ directly:



vs.



$\Rightarrow$ gives info. about uncertainty in $\vec{w}^*$ estimate

$\Rightarrow$ generally hard: Bayesian neural network

$\Rightarrow$ neat trick from stat. mech. that enables this (we will return to this later)