Main Idea - Assume $\Theta$ is drawn from a distribution

World $\rightarrow$ data depends on $\Theta$ $\rightarrow$ 'learn' $\Theta$

unknown parameters $\Theta$

'model' for what is unknown
Bayesian basics
let $X$ and $Y$ be discrete rvs taking values in $\mathbb{N}$. denote the joint pmf:

$$p_{XY}(x, y) = \mathbb{P}[X = x, Y = y]$$

**marginalization**: computing individual pmfs from joint pmfs as

$$p_X(x) = \sum_{y \in \mathbb{N}} p_{XY}(x, y) \quad p_Y(y) = \sum_{x \in \mathbb{N}} p_{XY}(x, y)$$

**conditioning**: pmf of $X$ given $Y = y$ (with $p_Y(y) > 0$) defined as:

$$\mathbb{P}[X = x | Y = y] \overset{\Delta}{=} p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)}$$

more generally, can define $\mathbb{P}[X \in A | Y \in B]$ for sets $A, B \in \mathbb{N}$

see also this visual demonstration
the basic ‘rules’ of Bayesian inference

let $X$ and $Y$ be discrete rvs taking values in $\mathbb{N}$, with joint pmf $p(x, y)$

**product rule**

$$h(x, y) = h(x) + h(y|x), H(x, y) = H(y) + H(x|y)$$

for $x, y \in \mathbb{N}$, we have:

$$p_{XY}(x, y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y)$$

**sum rule**

$$H(x) = H(y) + \sum_y p(y) H(x|y = y)$$

for $x \in \mathbb{N}$, we have:

$$p_X(x) = \sum_{y \in \mathbb{N}} p_{X|Y}(x|y)p_Y(y)$$

and most importantly!

**Bayes rule**

for any $x, y \in \mathbb{N}$, we have:

$$p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{\sum_{x \in \mathbb{N}} p_{Y|X}(y|x)p_X(x)}$$

see also [this video](#) for an intuitive take on Bayes rule
fundamental principle of Bayesian statistics

- assume the world arises via an underlying **generative model** $M$
- use **random variables** to model all unknown **parameters** $\theta$
- incorporate all that is known by conditioning on **data** $D$
- use Bayes rule to **update prior beliefs into posterior beliefs**

$$p(\theta|D,M) \propto p(\theta|M)p(D|\theta,M)$$

**Note**: Bayesian ML **DOES NOT** believe that the $\theta$ are random
- This is only for ‘convenience’
pros and cons

in praise of Bayes

- conceptually simple and easy to interpret
- works well with small sample sizes and overparametrized models
- can handle all questions of interest: no need for different estimators, hypothesis testing, etc.

why isn’t everybody Bayesian

- they need priors (subjectivity….) (but all methods are subjective….)
- they may be more computationally expensive: computing normalization constant and expectations, and updating priors, may be difficult

- Eg. MCMC (however, anytime you use a Bayesian ML method, you get much more info)
basics of Bayesian inference
given model $\mathcal{M}$ with parameters $\Theta$, and data $D$, we define:

- the **prior** $p(\Theta|\mathcal{M})$: what you believe before you see data
- the **posterior** $p(\Theta|D,\mathcal{M})$: what you believe after you see data
- the **marginal likelihood** or **evidence** $p(D|\mathcal{M})$: how probable is the data under our prior and model

these three are probability distributions; the next is not

- the **likelihood** $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M},\Theta)$: function of $\Theta$ summarizing data

**The likelihood principle** (main axiomatic basis for Bayesian ML)

given model $\mathcal{M}$, all evidence in data $D$ relevant to parameters $\Theta$ is contained in the likelihood function $\mathcal{L}(\Theta)$

this is not without controversy; see Wikipedia article
given model $\mathcal{M}$ with parameters $\Theta$, and data $D$, we define:

- the **prior** $p(\Theta|\mathcal{M})$: what you believe before you see data
- the **posterior** $p(\Theta|D,\mathcal{M})$: what you believe after you see data
- the **marginal likelihood** or **evidence** $p(D|\mathcal{M})$: how probable is the data under our prior and model
- the **likelihood**: $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M},\Theta)$: function of $\Theta$ summarizing the data

**the fundamental formula of Bayesian statistics**

$$
\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}
$$

$$
p(\theta|D,\mathcal{M}) = \frac{p(D|\theta,\mathcal{M}) p(\theta|\mathcal{M})}{p(D|\mathcal{M})}
$$

also see: **Sir David Spiegelhalter on Bayes vs. Fisher**

Most often (>90%)

$$
p(\theta|D,\mathcal{M}) \propto p(D|\theta,\mathcal{M}) p(\theta|\mathcal{M})
$$
Notes

- Likelihood, evidence are not distributions ($L(\theta)$ is just a fn of $\theta$)
  (which summarizes the data)

$P(\theta), P(\theta \mid D)$ are distributions over $\Theta$

- $L(\theta)$ is different for discrete vs continuous parameters $\Theta$
  - If $\theta$ discrete, $L(\theta \mid D) = P(\theta \mid D)$ (pmf)
  - If $\theta$ continuous, $L(\theta \mid D) = f(\theta \mid D)$ (pdf)

- The evidence is different for discrete vs continuous $D$
  - If $\theta$ discrete, evidence $= P(D \mid M)$
  - $\theta$ continuous, evidence $= f(D \mid M)$
example: the mystery Bernoulli rv

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$
- model $M$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution

Let $N_i = \# \text{ of } 1\text{s in } D$ and $N_0 = \# \text{ of } 0\text{s in } D$, then $N_i + N_0 = n$.

$$f(\theta) = \mathbb{P}(D| M, \theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i | M, \theta) = \theta^{N_i} (1-\theta)^{N_0}$$

Let $N = \# \text{ of } '1'\text{s in } \{X_1, X_2, \ldots, X_n\}$; what is $\mathbb{P}[H| M, \theta]$?

$$\mathbb{P}[N_i = k | \theta, M] = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$
the Bernoulli likelihood function

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$
- model $\mathcal{M}$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution

**likelihood:** $\mathcal{L}(\Theta) \triangleq p(D|\mathcal{M}, \theta)$: function of $\Theta$ summarizing the data

$$\mathcal{L}(\Theta) = \Theta^N \cdot (1-\Theta)^{N_0} \quad \theta \in [0, 1]$$

- Note - $\mathcal{L}(\Theta)$ is NOT a distribution, i.e.

$$\int \mathcal{L}(\Theta) d\Theta \neq 1$$
log-likelihood, sufficient statistics, MLE

- \( l(\theta) = \log L(\theta) = N_i \log \theta + N_0 \log (1-\theta) \) for Bernoulli

- \((N_i, N_0)\) are sufficient statistics
  - i.e., Given data \(D\), \(L(\theta)\) completely determined by \(N_i(D), N_0(D)\)

- MLE - max likelihood estimator
  \[
  \arg \max_{\theta \in [0,1]} \log L(\theta) = \arg \max_{\theta \in [0,1]} l(\theta) = \frac{N_i}{N_i + N_0}
  \]
Cromwell’s rule

How should we choose the prior?

The zeroth rule of Bayesian statistics

Never set $p(\theta|\mathcal{M}) = 0$ or $p(\theta|\mathcal{M}) = 1$ for any $\theta$

- Oliver Cromwell - ‘I beseech you, (supplication to higher authority), think it possible that you might be mistaken.’

- Connected to falsifiability

Also see:
- Jacob Bronowski on Cromwell’s Rule and the scientific method
- Richard Feynman on the scientific method (at Cornell!)
from where do we get a prior?

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$
- model $\mathcal{M}$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution

**option 1: from the ‘problem statement’**

Mackay example 2.6
- eleven urns labeled by $u \in \{0, 1, 2, \ldots, 10\}$, each containing ten balls
- urn $u$ contains $u$ red balls and $10 - u$ blue balls
- select urn $u$ uniformly at random and draw $n$ balls with replacement, obtaining $n_R$ red and $n - n_R$ blue balls

$$\Theta = \frac{i}{10} \quad \text{with prob} \quad \frac{1}{11} \quad \text{for each } i \in \{0, 1, \ldots, 10\}$$
from where do we get a prior

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$
- model $M$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution

option 2: the **maximum entropy** principle

choose $p(\theta|M)$ to be distribution with **maximum entropy** given $M$

we know $\theta \in [0, 1]$

\[ P \quad \text{If we know } \theta \in [0, 1], \text{ then one choice of prior } = \text{ Max Ent}(\{0, 1\}) \]

\[ = \text{ Unif}(\{0, 1\}) \]

\[ P \quad \text{If } \theta \in \mathbb{N}_+, E[\theta] = n \Rightarrow \text{Geom}(\frac{1}{\mu}) \]
from where do we get the prior, take 2

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0, 1\}^n$
- model $M$: $X_i$ are generated i.i.d. from a $Ber(\theta)$ distribution

**option 3: easy updates via conjugate priors**

- prior $p(\theta)$ is said to be conjugate to likelihood $p(D|\theta)$ if corresponding posterior $p(\theta|D)$ has same functional form as $p(\theta)$
- natural conjugate prior: $p(\theta)$ has same functional form as $p(D|\theta)$
- conjugate prior family: closed under Bayesian updating

**Note** - One obvious family = set of all distributions (not useful ...)

Want - 'smallest' family which is closed under Bayesian updating
Beta distribution

- $x \in [0, 1]$, parameters: $\Theta = (\alpha, \beta) \in \mathbb{R}^+$ (‘# ones’+1, ‘# zeros’+1)
- pdf: $p(x) \propto x^{\alpha-1} (1-x)^{\beta-1}$
- normalizing constant: $\frac{1}{B(\alpha, \beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$