PROBABILITY and UNCERTAINTY

ALEATORIC UNCERTAINTY

EPISTEMIC UNCERTAINTY

- FREQUENTIST INTERPETATION

  $P(X = i)$ probability of observing outcome i of an experiment

- BAYESIAN INTERPRETATION

  $P(X = i)$ belief of observing i

DISCRETE PROBABILITY DISTRIBUTIONS

$p(X=i) \in [0,1]$ $\quad \chi$ SET OF OUTCOMES

$$\sum_{i \in \chi} p(X=i) = 1$$

- RANDOM VARIABLE $x \in \chi$, $p(x=i)$

$p(x)$

X, Y RANDOM VARIABLES

$p(x,y)$ joint p.d. $\quad p(x=i, y=j)$

$p(x|y)$ conditioned probability of x given y (y has a fixed value)

$p(x)$ marginal p.d. of x

BERNOULLI $\quad x \in \{0,1\} \quad p(x) = \vartheta \in [0,1]$

COIN $\vartheta = \frac{1}{2}$

BINOMIAL DISTRIBUTION n THROWS

$$x \in \{0, \dots, n\} \qquad p(x=i) = \binom{n}{i} \vartheta^i (1-\vartheta)^{n-i}$$

LAWS OF PROBABILITY

- NORMALIZATION $\sum_{x \in X} p(x) = 1$ $\left(\sum_{i \in X} p(X = i) = 1 \right)$

- SUM RULE $p(x) = \sum_{y \in Y} p(x, y)$

- PRODUCT RULE $p(x, y) = p(x \mid y) p(y)$

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

- BAYES THEOREM

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{\underbrace{p(x)}}$$

$$p(x) = \sum_{y \in Y} p(x \mid y) p(y)$$

INDEPENDENCE

X and y are <u>independent</u>

$$P(x,y) = P(x)P(y)$$

$$\| \qquad \Rightarrow P(x|y) = P(x)$$

$$P(x|y)P(y)$$

CONTINUOUS PROBABILITY DISTRIBUTIONS

$\mathcal{X} \subseteq \mathbb{R}^n$ · $n \geq 1$

WE CONSIDER DISTRIBUTIONS HAVING A DENSITY (in $\mathbb{R}$)

$$p : \mathcal{X} \to \mathbb{R} \quad \text{s.t.} \quad p(x) \geq 0 \quad \text{and} \quad \int_{\mathcal{X}} p(x)\,dx = 1$$

$$p(\underbrace{x \in A}_{?}) = p(A) = \int_A p(x)\,dx$$

A is a nice subset (e.g. $A = [a,b] \subseteq \mathbb{R}$)

- $\int_{\mathcal{X}} p(x)\,dx = 1$
- $p(x) = \int_{\mathcal{Y}} p(x,y)\,dy$
- $p(x,y) = \underbrace{p(x|y)\,p(y)}$

EXPECTATIONS

$\mathcal{X} \ni x \qquad f : \mathcal{X} \to \mathbb{R}$

$$\mathbb{E}[f] = \mathbb{E}[f(x)] = \int_{\mathcal{X}} f(x)\, p(x)\, dx \quad \left(= \sum_{x \in \mathcal{X}} f(x) \cdot p(x) \right)$$

(discrete)

$x \in \{0,1\} \quad f(0) = -1000, \; f(1) = +100, \; p(x=1) = 0.75$

$$\mathbb{E}[f(x)] = -1000 \cdot (0.25) + 100\,(0.75) = -175$$

$\mu = \mathbb{E}[x]$ as mean

$\text{VAR} = \mathbb{E}\left[(x-\mu)^2\right]$ as variance

$$\text{COV}(x,y) = \mathbb{E}\left[(x-\mu_x)(y-\mu_y)\right]$$

$$\text{CORR}(x,y) = \frac{\text{cov}(x,y)}{\sqrt{\text{VAR}(x)\,\text{VAR}(y)}} \in [-1,1]$$

$$\mathbb{E}[f(x)] = \int_X f(x)\, p(x)\, dx$$

$X \rightarrow$ generate SAMPLES $x_1, \rightarrow x_N$ by using some sampling algorithm.

$$\mathbb{E}[f(x)] \approx \underbrace{\frac{1}{N} \sum_{i=1}^{N} f(x_i)}$$

$\downarrow N \rightarrow \infty$

$$\mathbb{E}[f(x)]$$

PROBABILITIES - THE FORMAL WAY

σ-ALGEBRA

$\mathcal{X}$ be a set, $\mathcal{S} \subseteq 2^{\mathcal{X}}$ is a σ-ALGEBRA

1) $\emptyset, \mathcal{X} \in \mathcal{S}$
2) $A \in \mathcal{S} \Rightarrow A^c \in \mathcal{S}$ $\quad (A^c = \mathcal{X} \setminus A)$
3) $A_n \in \mathcal{S} \Rightarrow \bigcup_n A_n \in \mathcal{S}, \quad n \in \mathbb{N}$

$(\mathcal{X}, \mathcal{S})$ IS A MEASURABLE SPACE.

$\mathcal{X} = \mathbb{R}^n$, $\mathcal{B} \subseteq 2^{\mathcal{X}}$ BOREL σ-ALGEBRA, the smallest σ-ALGEBRA containing open sets of the standard topology of $\mathbb{R}^n$

$(X, \mathcal{A})$, $(Y, \mathcal{B})$ measurable spaces

$f : (X, \mathcal{A}) \to (Y, \mathcal{B})$ is measurable

$f^{-1}(B) \in \mathcal{A}$, $B \in \mathcal{B}$

PROBABILITY MEASURE

$(X, \mathcal{A})$ measurable space. A p.m. on $(X, \mathcal{A})$ is

$p : \mathcal{A} \to [0,1]$ s.t.

$p(\emptyset) = 0$

$p(A^c) = 1 - p(A)$

if $A_n \in \mathcal{A}$ disjoint $(A_i \cap A_j = \emptyset)$

$p(\bigcup_n A_n) = \sum_n p(A_n)$

RANDOM VARIABLES

$(\Omega, \mathcal{S})$ — SAMPLE SPACE

$\omega \in \Omega$ state of the world

$P_\Omega(S), S \in \mathcal{S}$

$(\mathcal{X}, \mathcal{A})$ measurable spaces — SPACE of MEASURES

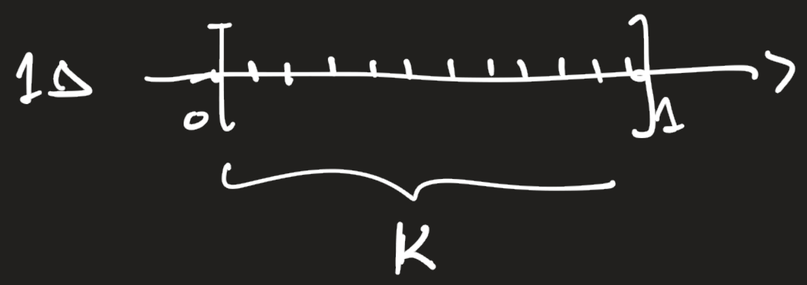
$x : \Omega \to \mathcal{X}$ measurable function

$A \in \mathcal{A} \Rightarrow x^{-1}(A) \in \mathcal{S}$

$P_x(A) = P_\Omega(x^{-1}(A))$

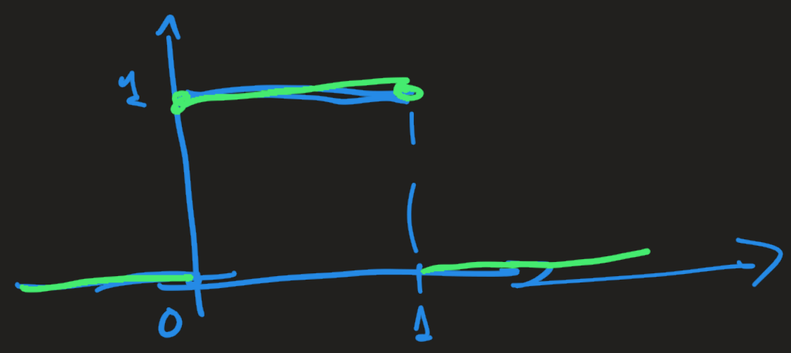
$P(x \in A) = \int_A p(x)\,dx$ — p: density of x

CURSE OF DIMENSIONALITY



UNIFORM Pd. ON $[0,1]$: $U(0,1)$

$u \sim U(0,1)$

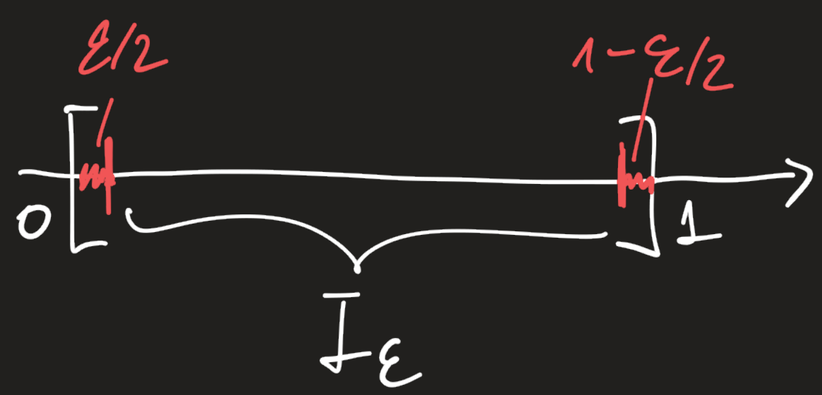


$$P(u \in [a,b]) = b-a = \int_a^b 1 \, du$$

$0 \le a \le b \le 1$

n dimension, a grid for $[0,1]^n$ will have K^n points

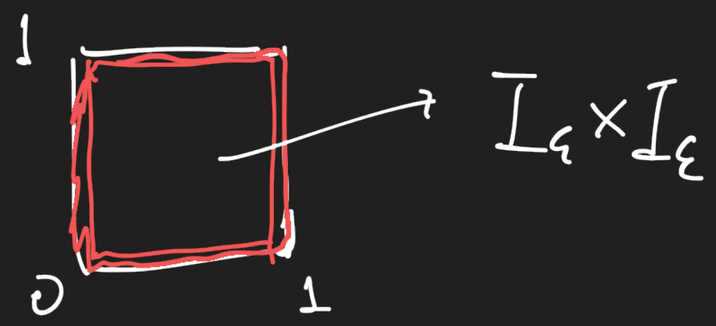
For n large, then K^n is out of reach



$$P(u \notin [\tfrac{\varepsilon}{2}, 1-\tfrac{\varepsilon}{2}]) = \varepsilon$$

$$P(u \in I_\varepsilon) = 1-\varepsilon$$

$$P(u \in I_\varepsilon^n) = (1-\varepsilon)^n \xrightarrow[n\to\infty]{} 0$$



$u = (u_1, u_2)$

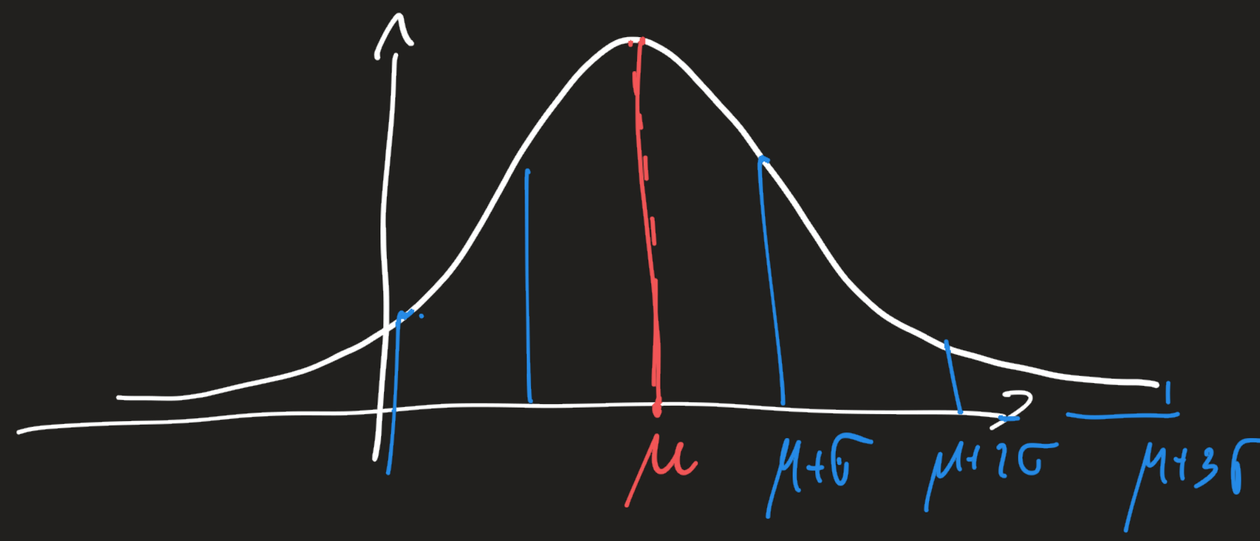
$P(u \in I_\varepsilon^2) = (1-\varepsilon)^2$

$u_j \sim U(0,1)$, u_j and u_i are indep. $j \neq i$

GAUSSIAN (NORMAL) DISTRIBUTION

$$p(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ: mean, σ^2: variance



$$p(\underline{x}|\underline{\mu},\underline{\Sigma}) = \mathcal{N}(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \underline{\Sigma}^{-1}(x-\mu)\right)$$

$\underline{x}$: n-dim vector, $\underline{\mu}$: n-dim vector, $\underline{\Sigma}$: $\Sigma_{ij} = \mathrm{cov}(X_i, X_j)$, $i,j = 1,\dots,n$

$\det(\Sigma)$

$\forall x.\ x^T \Sigma x \geq 0$

MIXTURES

$p_1(x), \quad \rightarrow p_K(x) \Rightarrow$ prob. densities

$x_1 \qquad\qquad x_K$

$\pi_1, \; - \; - \; , \pi_K \; , \; \sum \pi_i = 1 \; , \; \pi_i \in [0,1]$

$$p(x) = \sum_{i=1}^{K} \pi_i \, p_i(x) \qquad p_i(x) = \mathcal{N}(x \mid \mu_i, \Sigma_i)$$

$p(x)$ ↑

$p(x,z) \Rightarrow \quad p(x, z=i) = \underbrace{p(z=i)}_{\pi_i} \underbrace{p(x \mid z=i)}_{p_i(x)}$

$z \in \{1, \ldots, K\}$

$z \Rightarrow$ LATENT VARIABLE

$$p(x) = \sum_{i=1}^{K} p(x, z=i) = \sum_{i=1}^{K} p(z=i) \, p_i(x) = \sum_{i=1}^{K} \pi_i \, p_i(x)$$

$$= \sum_{z \in Z} p(x,z)$$

CONTINUOUS MIXTURES

$$p(x,z) = p(z)\,p(x|z)$$

$$p(x) = \int p(z)\,p(x|z)\,dz$$

$$p(z) = \text{GAMMA}(a,b)$$

$$p(x|z) = \mathcal{N}(x|\mu, z^{-1})$$

$p(x) \equiv$ t-student with mean μ

$\nu = 2a$ degrees of freedom

$\lambda = \frac{a}{b}$ scale

PROBABILISTIC INFERENCE

$H \in \{0,1\}$ - $H=1$ iff a person likes to work 14 hours per day

$F \in \{0,1\}$ - $F=1$ iff a person likes FRICO.

$H=1$ "$\Rightarrow$" $F=1$ (most likely)

$p(F \mid H=1) = 0.8$

$H=1$ is "rare" $\quad p(H=1) = 10^{-4}$

- If $F=1$ is common, how likely a $F=1$ is also $H=1$.

$p(F=1) = 0.4$

$$p(H=1 \mid F=1) = \frac{p(F=1 \mid H=1)\, p(H=1)}{p(F=1)} = \frac{0.8 \cdot 10^{-4}}{0.4} = 2 \cdot 10^{-4}$$

$$P(F=1) = 2 \cdot 10^{-4}$$

$$P(H=1 \mid F=1) = \frac{P(F=1 \mid H=1) \cdot P(H=1)}{P(F=1)} = \frac{0.8 \cdot 10^{-4}}{2 \cdot 10^{-4}} = 0.4$$

$$P(F=1) = P(F=1 \mid H=1) P(H=1) + P(F=1 \mid H=0) P(H=0)$$

$$P(x \mid y) = \frac{P(y \mid x) \cdot P(x)}{P(y)}$$

$$P(y) = \int P(y \mid x) P(x) dx$$

MAXIMUM LIKELIHOOD

$\underline{X}$: $x_1, \dots, x_N$ observation / data i.i.d

$p(x|\vartheta)$. there is a ϑ_{TRUE} that generates the data &

$$p(\underline{X}|\vartheta) = \prod_{i=1}^{N} p(x_i|\vartheta)$$ AS A FUNCTION OF ϑ.

M.L. IS FIND $\vartheta_{ML} = \arg\max_{\vartheta} p(\underline{X}|\vartheta)$

$$L(\vartheta) = \log p(\underline{X}|\vartheta) = \sum_{i=1}^{N} \log p(x_i|\vartheta)$$

$$\vartheta_{ML} = \arg\max_{\vartheta} L(\vartheta)$$

$x_1, \dots, x_N \in \{0,1\}$

$$p(x|\vartheta) = \text{Bernoulli}(x|\vartheta) \qquad ,\vartheta \in [0,1]$$
$$\overset{!}{=} \vartheta^x (1-\vartheta)^{1-x}$$

$$K = \sum_{i=1}^{N} x_i$$

$$\mathcal{L}(\vartheta) = \sum_{i=1}^{N} \log p(x_i|\vartheta) = \sum_{i=1}^{N} x_i \cdot \log(\vartheta) + (1-x_i)\log(1-\vartheta)$$
$$= K \cdot \log\vartheta + (N-K)\log(1-\vartheta)$$

$$0 \overset{!}{=} \frac{\partial \mathcal{L}(\vartheta)}{\partial \vartheta} = \frac{K}{\vartheta} - \frac{(N-K)}{1-\vartheta} = 0 \quad \Rightarrow \quad \vartheta = \frac{K}{N}$$

MAXIMUM A-POSTERIORI

$\underline{x}: x_1 \_ x_N$ iid

$p(x|\vartheta)$ - model

$p(\vartheta)$ - PRIOR

$$\underbrace{p(\vartheta|\underline{x})}_{\text{POSTERIOR ON } \vartheta \text{ GIVEN DATA}} = \frac{p(\underline{x}|\vartheta)\cdot p(\vartheta)}{\underbrace{p(\underline{x})}_{\text{MARGINAL LIKELIHOOD}}}$$

$$\vartheta_{MAP} = \arg\max_{\vartheta} p(\vartheta|\underline{x})$$

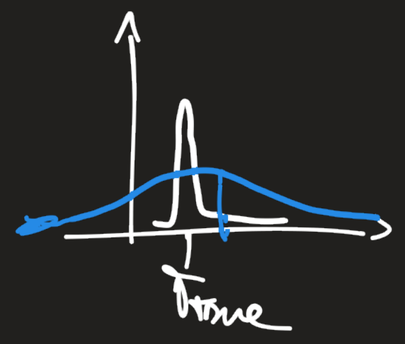
$$= \arg\max_{\vartheta} \log p(\underline{x}|\vartheta) + \log p(\vartheta)$$

BAYESIAN ESTIMATION

$\underline{x} : x_1 \to x_N$ iid

$p(x|\vartheta)$ LIKELIHOOD

$p(\vartheta)$ PRIOR



$$p|\vartheta(\underline{x}) = \frac{p(\underline{x}|\vartheta)\,p(\vartheta)}{p(\underline{x})}$$

POSTERIOR

$$p(\underline{x}) = \int p(x|\vartheta)\,p(\vartheta)\,d\vartheta$$

$$p(x|\underline{x}) = \int p(x|\vartheta)\,p(\vartheta|\underline{x})\,d\vartheta$$

PREDICTIVE DISTRIBUTION

$$\int p(x,\vartheta|\underline{x})\,d\vartheta = \int p(x|\vartheta)\,p(\vartheta|\underline{x})\,d\vartheta$$

$x_1 \dots x_N \in \{0,1\}$ $\quad p(x|\vartheta) = \text{Bernoulli}(\vartheta)$ $\quad \vartheta \in [0,1]$

$$p(\vartheta) = \text{Beta}(\vartheta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \vartheta^{\alpha-1}(1-\vartheta)^{\beta-1}$$

$$\mathbb{E}_{\text{Beta}(\alpha,\beta)}[\vartheta] = \frac{\alpha}{\alpha+\beta} \qquad K = \sum x_i$$

$$\log p(\vartheta|\underline{x}) = \underbrace{L(\vartheta) + \log B(\vartheta|\alpha,\beta)}_{C'} - \log p(\underline{x})$$

$$= C + K\log\vartheta + (N-K)\log(1-\vartheta) + (\alpha-1)\log\vartheta + (\beta-1)\log(1-\vartheta)$$

$$= C + (K+\alpha-1)\log\vartheta + (N-K+\beta-1)\log(1-\vartheta)$$

$$= \log \text{Beta}(\vartheta|K+\alpha, N-K+\beta)$$

Beta is Bernoulli CONJUGATE PRIOR

$$p(x_{\cdot}=1|\underline{x}) = \frac{K+\alpha}{N+\alpha+\beta}$$

INFORMATION THEORY PILLS - ENTROPY

$p(x)$

$-\log p(x)$ - SELF INFORMATION

$$H[p] = \mathbb{E}_p\left[-\log p(x)\right] = -\int p(x)\log p(x)\,dx$$

max is Normal RV.

$$H[p] = -\sum_i p(x_i)\log p(x_i)$$

max for uniform distr.

$\log K$.

KULLBACK LEIBLER DIVERGENCE

RELATIVE ENTROPY $\quad p, q$

$$KL[q \| p] = \int q(x) \log \frac{q(x)}{p(x)} dx = -H[q] - \underline{\mathbb{E}_q[\log p]}$$

$KL[q \| p] = 0 \iff q = p$

KL is a convex functional and $KL \geq 0$

$\rightsquigarrow$ p is fixed but unknown $\quad q = q_\theta$ can vary

What is the best q_θ that approx p.

$\theta^* = \underset{\theta}{\operatorname{argmin}} \; KL[q_\theta \| p]$ $\Big\{$ a VARIATIONAL INFERENCE

$I[x, y] = KL[p(x,y) \| p(x)\, p(y)]$ MUTUAL INFORMATION

$\underline{X} : X_1, \dots, X_N$

EMPIRICAL DISTRIBUTION $P_{emp}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(x = x_i)$

q

$$KL[P_{emp} \| q] = \underbrace{\mathbb{E}_{P_{emp}}[-\log q(x)]}_{-\frac{1}{N}\sum \log q(x_i)} - H[P_{emp}]$$

CROSS ENTROPY

If $q = q_\theta$ $\quad -\frac{1}{N} L(\theta) \rightsquigarrow$ MAXIMIZING $L(\theta)$

$\Updownarrow$

MINIMIZING K.L. divergence between P_{emp} and q_θ.