Problem Solving with AI Techniques
Refresher on Probability Theory

Paul Weng

UM-SJTU Joint Institute

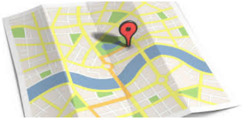
VE593, Fall 2018

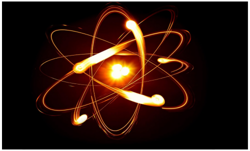
JOINT INSTITUTE
交大密西根学院

For more, see VE401 and VE501

Why is there Uncertainty?

- Knowledge is generally uncertain because
  - Full vs partial/non observable world
  - Observation comes from imperfect sensors
  - Data often imprecise, missing or contradictory
  - World is stochastic?
- Uncertainty can be handled with probability distribution
- Other uncertainty models, e.g., possibility theory, belief functions





Why Use Probability?

- Cox Theorem
  - If some "natural" postulates are satisfied, then belief is represented by probability

Why Use Probability?

- Cox Theorem
  - If some "natural" postulates are satisfied, then belief is represented by probability
- Dutch Book Argument: If rules of probability not respected, there exist a set of bets that guarantees to lose money (de Finetti). Example:

Proposition	Belief
a	.4
b	.3
$a \vee b$	.8

Why Use Probability?

- Cox Theorem
  - If some "natural" postulates are satisfied, then belief is represented by probability
- Dutch Book Argument: If rules of probability not respected, there exist a set of bets that guarantees to lose money (de Finetti). Example:

Proposition	Belief	Bet	Stakes
a	.4	a	4 to 6
b	.3	b	3 to 7
$a \vee b$	.8	$\neg(a \vee b)$	2 to 8

- Empirical Justification
  - Many success stories in gambling, finance, engineering, machine learning...

Interpretation of Probability

- Objective Probability
  - probability = objective property of objects like mass
- Frequentist Interpretation
  - probability = limit of observed frequences
- Subjective Interpretation
  - probability = beliefs of an agent

Probability

- Sample space/domain Ω
  Sample point/possible world/atomic event $\omega \in \Omega$
  Event $A \subseteq \Omega$
  e.g., Roll of 6-face die: $\Omega = \{1, 2, 3, 4, 5, 6\}$

Probability

- Sample space/domain Ω
  Sample point/possible world/atomic event $\omega \in \Omega$
  Event $A \subseteq \Omega$
  e.g., Roll of 6-face die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Probability $\mathbb{P} : A \subseteq \Omega \mapsto [0, 1]$
  e.g., $\mathbb{P}(\{1\}) = \frac{1}{6}$, $\mathbb{P}(\{2, 4\}) = \frac{1}{3}$, $\mathbb{P}(\{3, 5, 6\}) = \frac{1}{2}$

Probability

- Sample space/domain Ω
  Sample point/possible world/atomic event $\omega \in \Omega$
  Event $A \subseteq \Omega$
  e.g., Roll of 6-face die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Probability $\mathbb{P} : A \subseteq \Omega \mapsto [0, 1]$
  e.g., $\mathbb{P}(\{1\}) = \frac{1}{6}$, $\mathbb{P}(\{2, 4\}) = \frac{1}{3}$, $\mathbb{P}(\{3, 5, 6\}) = \frac{1}{2}$
- Kolmogorov Axioms: $\forall A, B \subseteq \Omega$
  - Nonnegativity $\mathbb{P}(A) \geq 0$
  - Normalization $\mathbb{P}(\Omega) = 1$
  - Additivity $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if $A \cap B = \{\,\}$

Probability

- Sample space/domain Ω
  Sample point/possible world/atomic event $\omega \in \Omega$
  Event $A \subseteq \Omega$
  e.g., Roll of 6-face die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Probability $\mathbb{P} : A \subseteq \Omega \mapsto [0, 1]$
  e.g., $\mathbb{P}(\{1\}) = \frac{1}{6}$, $\mathbb{P}(\{2, 4\}) = \frac{1}{3}$, $\mathbb{P}(\{3, 5, 6\}) = \frac{1}{2}$
- Kolmogorov Axioms: $\forall A, B \subseteq \Omega$
  - Nonnegativity $\mathbb{P}(A) \geq 0$
  - Normalization $\mathbb{P}(\Omega) = 1$
  - Additivity $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if $A \cap B = \{\,\}$
- Implications
  - $0 \leq \mathbb{P}(A) \leq 1$
  - $\mathbb{P}(\{\,\}) = 0$
  - $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
  - $\mathbb{P}(\Omega \backslash A) = 1 - \mathbb{P}(A)$
  - $\mathbb{P}(\cap_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i)$ if A_i's are disjoint

Random variable

- Random variable X: *measurable* function from Ω to some *measurable* space, e.g., $\mathbb{R}$ or $\{true, false\}$
  e.g., $X : \begin{cases} \{1,2,3,4,5,6\} \to \{1,2,3,4,5,6\} \\ x \mapsto x \end{cases}$, $Odd(\omega)$

Random variable

- Random variable X: *measurable* function from Ω to some *measurable* space, e.g., $\mathbb{R}$ or $\{true, false\}$
  e.g., $X : \begin{cases} \{1,2,3,4,5,6\} \rightarrow \{1,2,3,4,5,6\} \\ x \mapsto x \end{cases}$, $Odd(\omega)$
- For any random variable X, $\mathbb{P}$ induces a probability distribution, denoted $\mathbb{P}(X)$

$$\mathbb{P}(X = x_i) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x_i\}) = \sum_{\{\omega : X(\omega) = x_i\}} \mathbb{P}(\omega)$$

e.g., $\mathbb{P}(Odd = true) = \mathbb{P}(1) + \mathbb{P}(3) + \mathbb{P}(5) = 1/6 + 1/6 + 1/6 = 1/2$

Random variable

- Random variable X: *measurable* function from Ω to some *measurable* space, e.g., $\mathbb{R}$ or $\{true, false\}$
  e.g., $X : \left\{ \begin{array}{l} \{1,2,3,4,5,6\} \to \{1,2,3,4,5,6\} \\ x \mapsto x \end{array} \right.$, $Odd(\omega)$
- For any random variable X, $\mathbb{P}$ induces a probability distribution, denoted $\mathbb{P}(X)$

$$\mathbb{P}(X = x_i) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x_i\}) = \sum_{\{\omega : X(\omega) = x_i\}} \mathbb{P}(\omega)$$

  e.g., $\mathbb{P}(Odd = true) = \mathbb{P}(1) + \mathbb{P}(3) + \mathbb{P}(5) = 1/6 + 1/6 + 1/6 = 1/2$
- For finite space, think of $\mathbb{P}(X)$ as a table

Random variable

- Random variable X: *measurable* function from Ω to some *measurable* space, e.g., $\mathbb{R}$ or $\{true, false\}$
  e.g., $X : \begin{cases} \{1,2,3,4,5,6\} \to \{1,2,3,4,5,6\} \\ x \mapsto x \end{cases}$, $Odd(\omega)$
- For any random variable X, $\mathbb{P}$ induces a probability distribution, denoted $\mathbb{P}(X)$

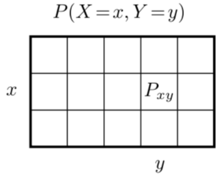
$$\mathbb{P}(X = x_i) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x_i\}) = \sum_{\{\omega : X(\omega) = x_i\}} \mathbb{P}(\omega)$$

  e.g., $\mathbb{P}(Odd = true) = \mathbb{P}(1) + \mathbb{P}(3) + \mathbb{P}(5) = 1/6 + 1/6 + 1/6 = 1/2$
- For finite space, think of $\mathbb{P}(X)$ as a table
- Notation $\sum_X \mathbb{P}(X) = \sum_x \mathbb{P}(X = x)$

Joint Distribution

Assume we have two random variables X and Y

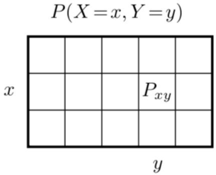
- Joint $\mathbb{P}(X, Y)$



Joint Distribution

Assume we have two random variables X and Y

- Joint $\mathbb{P}(X, Y)$
- Marginal $\mathbb{P}(X) = \sum_Y \mathbb{P}(X, Y)$

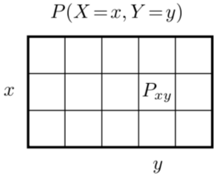


Joint Distribution

Assume we have two random variables X and Y

- Joint $\mathbb{P}(X, Y)$
- Marginal $\mathbb{P}(X) = \sum_Y \mathbb{P}(X, Y)$
- Conditional $\mathbb{P}(X \mid Y) = \frac{\mathbb{P}(X,Y)}{\mathbb{P}(Y)}$

  What is the dimension of $\mathbb{P}(X \mid Y)$?
  What is the value of $\sum_X \mathbb{P}(X \mid Y)$?



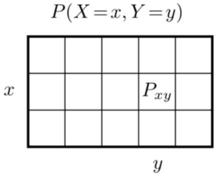
Joint Distribution

Assume we have two random variables X and Y

- Joint $\mathbb{P}(X, Y)$
- Marginal $\mathbb{P}(X) = \sum_Y \mathbb{P}(X, Y)$
- Conditional $\mathbb{P}(X \mid Y) = \frac{\mathbb{P}(X,Y)}{\mathbb{P}(Y)}$

  What is the dimension of $\mathbb{P}(X \mid Y)$?
  What is the value of $\sum_X \mathbb{P}(X \mid Y)$?
- X is independent of Y iff $\mathbb{P}(X \mid Y) = \mathbb{P}(X)$

  What does it mean for $\mathbb{P}(X \mid Y)$ and $\mathbb{P}(Y \mid X)$?



Important Rules

- Product rule

$$\mathbb{P}(X, Y) = \mathbb{P}(X \mid Y)\mathbb{P}(Y) = \mathbb{P}(Y \mid X)\mathbb{P}(X)$$

- Bayes' rule

$$\mathbb{P}(X \mid Y) = \frac{\mathbb{P}(Y \mid X)\mathbb{P}(X)}{\mathbb{P}(Y)}$$

$\mathbb{P}(X \mid Y)$ posterior	$\mathbb{P}(X)$ prior
$\mathbb{P}(Y \mid X)$ likelihood	$\mathbb{P}(Y)$ normalization

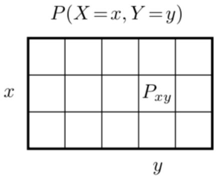
- Extended form

$$\mathbb{P}(X \mid Y) = \frac{\mathbb{P}(Y \mid X)\mathbb{P}(X)}{\sum_X \mathbb{P}(Y \mid X)\mathbb{P}(X)}$$

Joint Distribution: General Case

Assume we have n random variables $X_{1:n}$

- Joint $\mathbb{P}(X_{1:n})$
- Marginal $\mathbb{P}(X_1) = \sum_{X_{2:n}} \mathbb{P}(X_{1:n})$
- Conditional $\mathbb{P}(X_1 \mid X_{2:n}) = \frac{\mathbb{P}(X_{1:n})}{\mathbb{P}(X_{2:n})}$
- X is independent of Y given Z iff $\mathbb{P}(X \mid Y, Z) = \mathbb{P}(X \mid Z)$



$\Downarrow$

n-th order tensor

Expectation

- Discrete case: $\mathbb{E}[X] = \sum_x x\mathbb{P}(X = x)$
- Continuous case: $\mathbb{E}[X] = \int_x xp(x)dx$

- More generally, $\mathbb{E}[f(X)]$ for some function f
- $\mathbb{E}[\boldsymbol{X}]$ if $\boldsymbol{X} \in \mathbb{R}^n$

- Operator $\mathbb{E}$ is linear:
  - $\mathbb{E}(\lambda X) = \lambda\mathbb{E}(X)$
  - $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

Variance

- $\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$

- Covariance $cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
- If X, Y are independent, then $cov(X, Y) = 0$. But opposite generally not true!

- Useful identities:
  - $\mathbb{V}(\lambda X) = \lambda^2 V(X)$
  - $\mathbb{V}(X + Y) = \mathbb{V}(X) + 2cov(X, Y) + \mathbb{V}(Y)$

- If $\boldsymbol{X} \in \mathbb{R}^n$, $\mathbb{V}[\boldsymbol{X}] = \mathbb{E}[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{X} - E[\boldsymbol{X}])^\intercal]$

Bernoulli & Binomial Distribution

- Random variable $X \in \{0, 1\}$ follows a Bernoulli distribution $Bern(p)$

$$\mathbb{P}(X = 1 \mid p) = p, \mathbb{P}(X = 0 \mid p) = 1 - p, \quad \mathbb{P}(X = x \mid p) = p^x (1-p)^{1-x}$$

Bernoulli & Binomial Distribution

- Random variable $X \in \{0, 1\}$ follows a Bernoulli distribution $Bern(p)$

  $\mathbb{P}(X = 1 \mid p) = p, \mathbb{P}(X = 0 \mid p) = 1 - p, \quad \mathbb{P}(X = x \mid p) = p^x(1-p)^{1-x}$

- Dataset of i.i.d. random variables $D = (X_1, X_2, \dots, X_n)$ where $X_i \sim Bern(p)$

$$\mathbb{P}(D = (x_1, \dots, x_n) \mid p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$\arg\max_p \mathbb{P}(D \mid p) = \arg\max_p \sum_{i=1}^{n} x_i \ln p + (1 - x_i) \ln(1 - p) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Bernoulli & Binomial Distribution

- Random variable $X \in \{0, 1\}$ follows a Bernoulli distribution $Bern(p)$

$$\mathbb{P}(X = 1 \mid p) = p, \mathbb{P}(X = 0 \mid p) = 1 - p, \quad \mathbb{P}(X = x \mid p) = p^x (1-p)^{1-x}$$

- Dataset of i.i.d. random variables $D = (X_1, X_2, \dots, X_n)$ where $X_i \sim Bern(p)$

$$\mathbb{P}(D = (x_1, \dots, x_n) \mid p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

$$\arg\max_p \mathbb{P}(D \mid p) = \arg\max_p \sum_{i=1}^{n} x_i \ln p + (1 - x_i) \ln(1-p) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Random variable $M = \sum_{i=1}^{n} X_i \sim$ Binomial distribution $Bin(n, p)$

$$\mathbb{P}(M = m \mid n, p) = \binom{n}{m} p^m (1-p)^{n-m}, \quad \binom{n}{m} = \frac{n!}{(n-m)!m!}$$

How to Express Uncertainty over a Bernoulli Parameter p?

- Beta distribution $Beta(\alpha, \beta)$ with $\alpha, \beta > 0$ = distribution over $[0, 1]$

$$Beta(p \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad \text{with mean } \tfrac{\alpha}{\alpha+\beta}$$

How to Express Uncertainty over a Bernoulli Parameter p?

- Beta distribution $Beta(\alpha, \beta)$ with $\alpha, \beta > 0$ = distribution over $[0, 1]$

$$Beta(p \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad \text{with mean } \tfrac{\alpha}{\alpha+\beta}$$

- It can be used to represent belief about unknown p:
  $\mathbb{P}(p) = Beta(p \mid \alpha, \beta)$

How to Express Uncertainty over a Bernoulli Parameter p?

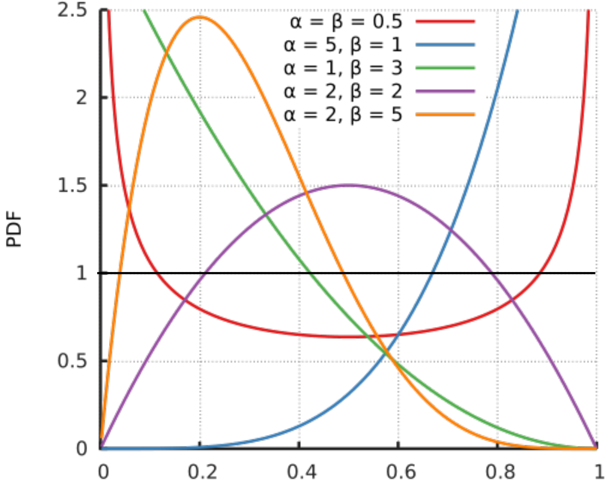
- Beta distribution $Beta(\alpha, \beta)$ with $\alpha, \beta > 0$ = distribution over $[0, 1]$

$$Beta(p \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad \text{with mean } \tfrac{\alpha}{\alpha+\beta}$$

- It can be used to represent belief about unknown p: $\mathbb{P}(p) = Beta(p \mid \alpha, \beta)$
- After observing $D = (x_1, \ldots, x_n)$, with counts $a = \sum_i x_i$ and $b = \sum_i (1 - x_i)$, the belief about p can be updated:

$$\begin{aligned}
\mathbb{P}(p \mid D) &= \frac{\mathbb{P}(D \mid p)\mathbb{P}(p)}{\mathbb{P}(D)} \propto Bin(D \mid p) Beta(p \mid \alpha, \beta) \\
&\propto p^a (1-p)^b p^{\alpha-1}(1-p)^{\beta-1} = p^{\alpha-1+a}(1-p)^{\beta-1+b} \\
&= Beta(\alpha + a, \beta + b)
\end{aligned}$$

Examples of Beta Distribution



Categorical & Multinomial Distribution

- Random variable $X \in \{1, 2, \dots K\}$ follows a categorical distribution $Cat(\boldsymbol{p})$ with $\boldsymbol{p} = (p_1, \dots, p_K)$ and $\sum_k p_k = 1$

$$\mathbb{P}(X = k \mid \boldsymbol{p}) = p_k$$

Categorical & Multinomial Distribution

- Random variable $X \in \{1, 2, \dots K\}$ follows a categorical distribution $Cat(\boldsymbol{p})$ with $\boldsymbol{p} = (p_1, \dots, p_K)$ and $\sum_k p_k = 1$

$$\mathbb{P}(X = k \mid \boldsymbol{p}) = p_k$$

- Dataset of i.i.d. random variables $D = (X_1, \dots, X_n)$ with $X_i \sim Cat(\boldsymbol{p})$

$$\mathbb{P}(D = (x_1, \dots, x_n) \mid \boldsymbol{p}) = \prod_{i=1}^{n} p_{x_i} = \prod_{i=1}^{n} \prod_{k=1}^{K} p_k^{[x_i = k]} = \prod_{k=1}^{K} p_k^{m_k}$$

where $m_k = \sum_{i=1}^{n} [x_i = k]$.

$$\arg\max_{\boldsymbol{p}} \mathbb{P}(D \mid \boldsymbol{p}) = \frac{1}{n}(m_1, m_2, \dots, m_K)$$

Categorical & Multinomial Distribution

- Random variable $X \in \{1, 2, \dots K\}$ follows a categorical distribution $Cat(\boldsymbol{p})$ with $\boldsymbol{p} = (p_1, \dots, p_K)$ and $\sum_k p_k = 1$

$$\mathbb{P}(X = k \mid \boldsymbol{p}) = p_k$$

- Dataset of i.i.d. random variables $D = (X_1, \dots, X_n)$ with $X_i \sim Cat(\boldsymbol{p})$

$$\mathbb{P}(D = (x_1, \dots, x_n) \mid \boldsymbol{p}) = \prod_{i=1}^{n} p_{x_i} = \prod_{i=1}^{n} \prod_{k=1}^{K} p_k^{[x_i = k]} = \prod_{k=1}^{K} p_k^{m_k}$$

where $m_k = \sum_{i=1}^{n} [x_i = k]$.

$$\arg\max_{\boldsymbol{p}} \mathbb{P}(D \mid \boldsymbol{p}) = \frac{1}{n}(m_1, m_2, \dots, m_K)$$

- Random variable $\boldsymbol{M} = (\sum_{i=1}^{n} [X_i = k])_{k=1,\dots K}$ follows a multinomial distribution $Mult(n, \boldsymbol{p})$

$$\mathbb{P}(\boldsymbol{M} = (m_1, \dots, m_K) \mid n, \boldsymbol{p}) \propto \prod_{k=1}^{K} p_k^{m_k}$$

How to Express Uncertainty over Multinomial Parameter $\boldsymbol{p}$?

- Dirichlet distribution $Dir(\boldsymbol{\alpha})$ with $\alpha_k > 0$ = distribution over $(K-1)$-simplex: $\{\boldsymbol{p} \mid \boldsymbol{p} \geq 0, \sum_k p_k = 1\}$

$$Dir(\boldsymbol{p} \mid \boldsymbol{\alpha}) \propto \prod_{k=1}^{K} p_k^{\alpha_k - 1} \quad \text{with mean } \left(\frac{\alpha_k}{\sum_j \alpha_j}\right)_{k=1,\ldots,K}$$

How to Express Uncertainty over Multinomial Parameter $\boldsymbol{p}$?

- Dirichlet distribution $Dir(\boldsymbol{\alpha})$ with $\alpha_k > 0$ = distribution over $(K-1)$-simplex: $\{\boldsymbol{p} \mid \boldsymbol{p} \geq 0, \sum_k p_k = 1\}$

$$Dir(\boldsymbol{p} \mid \boldsymbol{\alpha}) \propto \prod_{k=1}^{K} p_k^{\alpha_k - 1} \quad \text{with mean } \Big(\frac{\alpha_k}{\sum_j \alpha_j}\Big)_{k=1,\dots,K}$$

- It can be used to represent belief about unknown $\boldsymbol{p}$: $\mathbb{P}(\boldsymbol{p}) = Dir(\boldsymbol{p} \mid \boldsymbol{\alpha})$

How to Express Uncertainty over Multinomial Parameter $\boldsymbol{p}$?

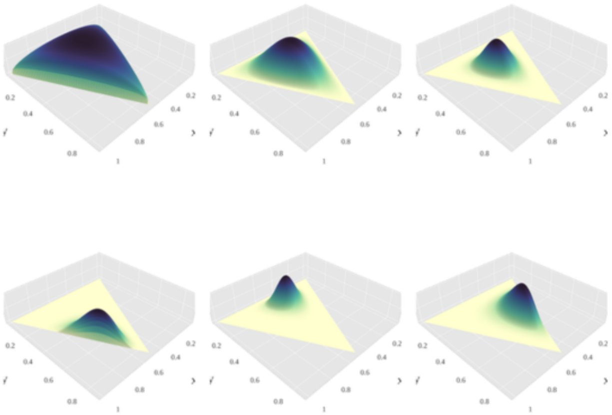
- Dirichlet distribution $Dir(\boldsymbol{\alpha})$ with $\alpha_k > 0 =$ distribution over $(K-1)$-simplex: $\{\boldsymbol{p} \mid \boldsymbol{p} \geq 0, \sum_k p_k = 1\}$

$$Dir(\boldsymbol{p} \mid \boldsymbol{\alpha}) \propto \prod_{k=1}^{K} p_k^{\alpha_k - 1} \quad \text{with mean } \Big(\frac{\alpha_k}{\sum_j \alpha_j}\Big)_{k=1,\dots,K}$$

- It can be used to represent belief about unknown $\boldsymbol{p}$: $\mathbb{P}(\boldsymbol{p}) = Dir(\boldsymbol{p} \mid \boldsymbol{\alpha})$
- After observing $D = \{x_1, \dots, x_n\}$, with counts $a_k = \sum_i [x_i = k]$, the belief about p can be updated:

$$\begin{aligned}
\mathbb{P}(\boldsymbol{p} \mid D) &= \frac{\mathbb{P}(D \mid \boldsymbol{p})\mathbb{P}(\boldsymbol{p})}{\mathbb{P}(D)} \propto Mult(D \mid \boldsymbol{p}) Dir(\boldsymbol{p} \mid \boldsymbol{\alpha}) \\
&\propto \prod_{k=1}^{K} p_k^{a_k} \prod_{k=1}^{K} p_k^{\alpha_k - 1} = \prod_{k=1}^{K} p_k^{\alpha_k - 1 + a_k} \\
&= Dir(\boldsymbol{\alpha} + \boldsymbol{a})
\end{aligned}$$

Examples of Dirichlet Distribution



Other Discrete Distributions

- Uniform discrete distribution
- Geometric distribution: How many Bernoulli trials before a success?
- Negative binomial distribution: How many successes in a sequence of Bernoulli trials with a fixed number of failures?
- Poisson distribution: How many successes in a duration of time if they occur at known constant rate?

Distributions over Continuous Domains

- Probability for $X \in \mathbb{R}$ determined by probability density function $p(x) \in [0, \infty)$:

$$\mathbb{P}(X \in [a, b]) = \int_a^b p(x)dx \in [0, 1]$$

- Cumulative probability distribution:

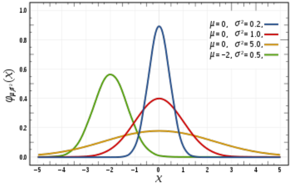
$$F(y) = \mathbb{P}(X \leq y) = \int_{-\infty}^{y} p(x)dx$$

- Note: for continuous probability distribution, $\mathbb{P}(X = x) = 0$

Gaussian (or Normal) Distribution

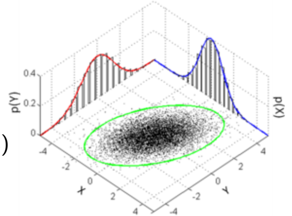
- on $\mathbb{R}$:

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



- on $\mathbb{R}^n$:

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$



Why are Gaussian Distributions Important?

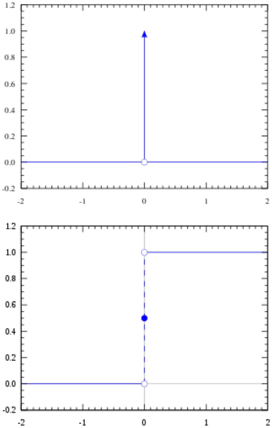
- Central limit theorem: Averages of n i.i.d random variables (with mean μ and variance σ^2) $\sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
- If only mean and variance are known, it is the distribution that maximizes *entropy*
- It makes math simpler, e.g., weighted sum of independent Gaussian r.v.s is also Gaussian

Dirac Distribution

- Dirac distribution $\delta(x) = 0$ except at $x = 0$ such that

$$\int \delta(x)dx = 1$$

- $\delta(x) = \frac{\delta}{\delta x}H(x)$ where $H(x) = [x \geq 0]$, Heaviside step function
- Limit of $\mathcal{N}(0, \frac{\sigma^2}{n})$ as $n \to \infty$
- Can represent certainty



Other Continuous Distributions

- Beta and Dirichlet distributions
- Continuous uniform distribution over a compact set (e.g., interval)
- Exponential distribution: How long before an event happens if it occurs at some known rate?
- Logistic distribution: distribution whose CDF is the logistic function $\frac{1}{1+e^{-\frac{x-\mu}{s}}}$
- χ^2 distribution: distribution of sum of n squared Gaussian

For more, see VE550

Entropy

- Neg-log of a distribution ($-\log p(x)$) reflects something like "error":
  - neg-log of Gaussian $\leftrightarrow$ squared error
  - neg-log of likelihood $\leftrightarrow$ prediction error
- Term $-\log p(x)$ is "optimal" coding length you should assign to symbol x. This will minimize the expected length of an encoding:

$$H(p) = \int_x p(x)\big(-\log p(x)\big)dx \geq 0$$

- Entropy $H(X) = \mathbb{E}[-\log p(X)] =$ measure of uncertainty, or lack of information, we have about X
- Note: Uniform distribution has highest entropy and *Dirac* distribution has lowest entropy

Relative Entropy or Kullback-Leibler Divergence

- Assume distribution $q(x)$ used to decide on coding length of symbols drawn from $p(x)$. Expected length of encoding is given by *cross-entropy*:

$$\int_x p(x)\big(-\log q(x)\big)\,dx \geq H(p)$$

- Difference

$$D(p||q) = \int_x p(x)\Big(\log \frac{p(x)}{q(x)}\Big)\,dx \geq 0$$

is called relative entropy or Kullback-Leibler divergence

- Note: Although not a distance, it can be used to measure how different two distributions are