Probability and Information Theory

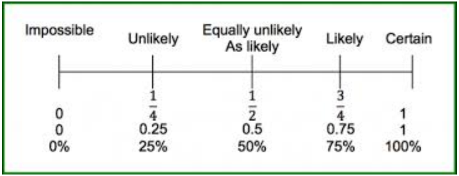
Changyou Chen

Department of Computer Science and Engineering
Universitpy at Buffalo, SUNY
changyou@buffalo.edu

Feburary 7, 2019

What is Probability?

1. Probability is a measure of the likelihood of an event happening.
2. We measure probability on a scale from 0 to 1:
   - A probability of 1 indicates the event is certain to happen.
   - A probability of 0 indicates the event is certainly not happen.



Probability Mass Function

1. Define for discrete random variables; denote as P.
2. The domain of P is the set of all possible states (values) of x.
3. $\forall x, 0 \leq P(x) \leq 1$:
   - An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring.
4. $\sum_x P(x) = 1$:
   - Refereed to as being normalized. Without this property, we could obtain probabilities greater than one or unbounded.

Example: uniform distribution: $P(\mathbf{x} = x_i) = \frac{1}{|\mathbf{x}|}$, where $|\,\mathbf{x}\,|$ is the number of values $\mathbf{x}$ could take.

Probability Density Function

1. Defined for continuous random variables; Denote as p.
2. The domain of p must be the set of all possible states (values) of x.
3. $\forall x, p(x) \geq 0$. Note that we do not require $p(x) \leq 1$.
4. As long as $\int p(x) \mathrm{d}x = 1$.

Example: uniform distribution: $u(\mathbf{x}; a, b) = \frac{1}{b-a}$

Computing Marginal Probability with the Sum Rule

Given joint distribution $P(\mathbf{x}, \mathbf{y})$ (for discrete random variables) or $p(\mathbf{x}, \mathbf{y})$ (for continuous random variables):

- the probability of the events ($\mathbf{x} = x$, *e.g.*, raining) and ($\mathbf{y} = y$, *e.g.*, running) jointly happen.

$$\forall x, P(\mathbf{x} = x) = \sum_y P(\mathbf{x} = x, \mathbf{y} = y)$$

$$p(\mathbf{x} = x) = \int p(\mathbf{x} = x, \mathbf{y} = y)\mathrm{d}y$$

Conditional Probability

- The probability of an event ($\mathbf{y} = y$, *e.g.*, running) given the other event ($\mathbf{x} = x$, *e.g.*, raining) happened.
- Applies for both discrete and continuous random variables.

$$P(\mathbf{y} = y | \mathbf{x} = x) = \frac{P(\mathbf{y} = y, \mathbf{x} = x)}{P(\mathbf{x} = x)}$$

Chain Rule of Probability

- Applies for both discrete and continuous random variables.

$$P(x^{(1)}, \cdots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^{n} P(x^{(i)} | x^{(1)}, \cdots, x^{(i-1)})$$

Independence/Conditional Independence

- Applies for both discrete and continuous random variables.
- Independence:

$$\forall x \in \mathbf{x}, y \in \mathbf{y}, p(\mathbf{x} = x, \mathbf{y} = y) = p(\mathbf{x} = x)p(\mathbf{y} = y)$$

- Conditional independence:

$$\begin{aligned} &\forall x \in \mathbf{x}, y \in \mathbf{y}, z \in \mathbf{z}, \\ &p(\mathbf{x} = x, \mathbf{y} = y | \mathbf{z} = z) = p(\mathbf{x} = x | \mathbf{z} = z)p(\mathbf{y} = y | \mathbf{z} = z) \end{aligned}$$

Expectation

- Average value of a function under a probability distribution.

$$\mathbb{E}_{x \sim P}[f(x)] = \sum_x P(x) f(x)$$
$$\mathbb{E}_{x \sim p}[f(x)] = \int p(x) f(x) \mathrm{d}x$$

linearity of expectations:

$$\mathbb{E}_x[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_x[f(x)] + \beta \mathbb{E}_x[g(x)]$$

Variance and Covariance

- Variance: how fluctuant a function is under a probability distribution.
- Covariance: how correlated two function is under a probability distribution.

$$\begin{aligned}
\mathrm{Var}\left(f(x)\right) &= \mathbb{E}_x\left[\left(f(x) - \mathbb{E}[f(x)]\right)^2\right] \\
\mathrm{Cov}\left(f(x), g(y)\right) &= \mathbb{E}_{(x,y)}\left[\left(f(x) - \mathbb{E}_x[f(x)]\right)\left(g(y) - \mathbb{E}_y[g(y)]\right)\right] \\
\text{discrete} \rightarrow &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(f(x_i) - \mathbb{E}[f(x)]\right)\big(g(y_j) - \mathbb{E}[g(y_j)]\big)
\end{aligned}$$

Covariance matrix:

$$\mathrm{Cov}(\mathbf{x})_{ij} = \mathrm{Cov}(x_i, x_j)$$

Bernoulli Distribution

- Binary random variables.

$$
\begin{aligned}
P(\mathbf{x} = 1) &= \phi \\
P(\mathbf{x} = 0) &= 1 - \phi \\
P(\mathbf{x} = x) &= \phi^{x}(1 - \phi)^{1-x}, \quad x \in \{0, 1\} \\
\mathbb{E}[\mathbf{x}] &= \phi \\
\mathrm{Var}(\mathbf{x}) &= \phi(1 - \phi)
\end{aligned}
$$

Gaussian Distribution

- Parameterized by mean μ and variance σ^2:

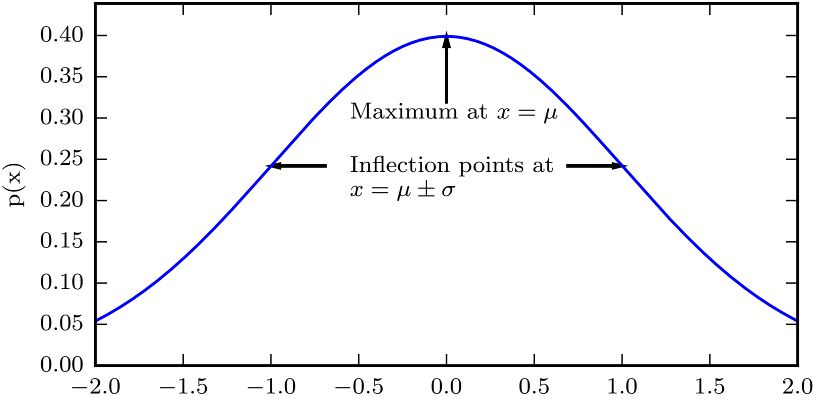
$$p(x;\mu,\sigma^2) \triangleq \mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

$$\mathbb{E}[x] = \mu, \quad \text{Var}(x) = \sigma^2$$

- Parameterized by mean μ and precision $\beta \triangleq \frac{1}{\sigma^2}$:

$$p(x;\mu,\beta) \triangleq \mathcal{N}(x;\mu,\beta^{-1}) = \sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{\beta}{2}(x-\mu)^2\right)$$

Gaussian Distribution



Multivariate Gaussian Distribution

- Parameterized by mean $\boldsymbol{\mu}$ and covariance matrix Σ:

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

$$\mathbb{E}\left[\mathbf{x}\right] = \boldsymbol{\mu}, \quad \mathrm{Cov}(\mathbf{x}) = \Sigma$$

- Parameterized by mean $\boldsymbol{\mu}$ and precision $\beta = \Sigma^{-1}$:

$$p(\mathbf{x}; \boldsymbol{\mu}, \beta) \triangleq \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \beta^{-1}) = \sqrt{\frac{\det(\beta)}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \beta (\mathbf{x} - \mu)\right)$$

More Distributions

- Exponential:

$$p(x;\lambda) = \lambda \mathbf{1}_{x\geq 0} \exp\left(-\lambda x\right)$$

- Laplace:

$$p(x;\mu,\gamma) = \frac{1}{2\lambda} \exp\left(-\frac{|x-\mu|}{\gamma}\right)$$

- Dirac:

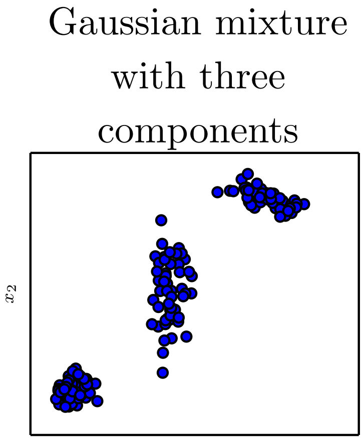
$$p(x) = \delta(x-\mu)$$

- Empirical Distribution:

$$\hat{p}(x) = \frac{1}{m}\sum_{i=1}^{m} \delta(x - x^{(i)})$$

Mixture Distributions

$$p(x) = \sum_i \underbrace{P(c = i)}_{\text{weights}} \underbrace{p(x|c = i)}_{\text{conditional distribution}}$$



Bayes' Rule: Learning from Data

- Let $\mathcal{D}$ be a given data set; $\boldsymbol{\theta}$ be the model parameter

$$p(\boldsymbol{\theta} \,|\mathcal{D}) = \frac{p(\mathcal{D}, \boldsymbol{\theta})}{p(\mathcal{D})} = \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\,\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathcal{D}|\,\boldsymbol{\theta})\mathrm{d}\,\boldsymbol{\theta}} = \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\,\boldsymbol{\theta})}{p(\mathcal{D})}\,.$$

- $p(\boldsymbol{\theta})$: prior distribution of $\boldsymbol{\theta}$
- $p(\mathcal{D}|\,\boldsymbol{\theta})$: likelihood of $\boldsymbol{\theta}$ on data
- $p(\boldsymbol{\theta} \,|\mathcal{D})$: posterior distribution
- $p(\mathcal{D}, \boldsymbol{\theta})$: joint distribution of data and model
- $p(\mathcal{D})$: marginal likelihood

Change of Random Variables

What is the probability distribution of **x** given the probability of **y** and a deterministic mapping $g(\cdot)$ from **x** to **y**?

$$p_x(x) = p_y(g(x)) \left| \det \left(\frac{\partial g(x)}{\partial x} \right) \right|$$

Change of Random Variables

What is the probability distribution of **x** given the probability of **y** and a deterministic mapping $g(\cdot)$ from **x** to **y**?

$$p_x(x) = p_y(g(x)) \left|\det\left(\frac{\partial g(x)}{\partial x}\right)\right|$$

Fundamental rule for normalizing flow based inference in variational autoencoder?

Information Theory

- Information:

$$I(x) = -\log P(x)$$

- Entropy (expected information):

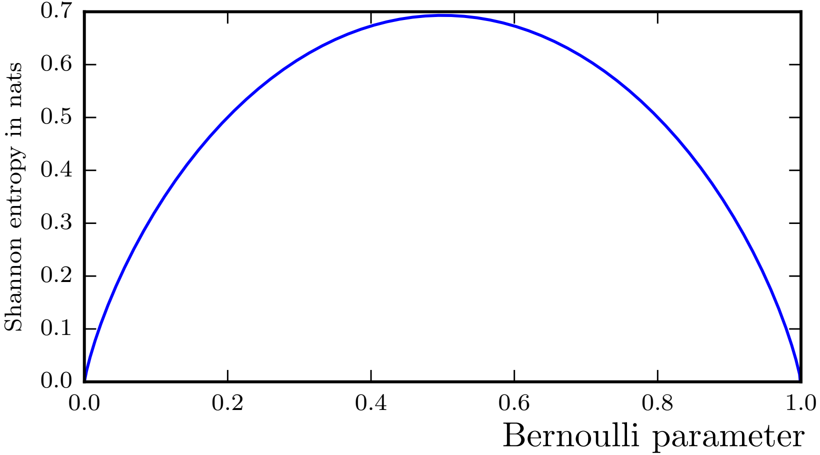
$$H(x) = \mathbb{E}_{\mathbf{x}\sim p}[I(x)] = -\mathbb{E}_{\mathbf{x}\sim p}[\log p(x)]$$

- KL divergence:

$$D_{KL}(P\|Q) = \mathbb{E}_{x\sim p}\left[\log\frac{p(x)}{Q(x)}\right] = \mathbb{E}_{x\sim p}[\log p(x) - \log Q(x)]$$

Entropy of a Bernoulli Random Variable

$$P(\mathbf{x} = 1) = \phi, \qquad P(\mathbf{x} = 0) = 1 - \phi$$



Asymmetric KL Divergence

- What does the optimal value look like under the constraints that p is a two-mode distribution and q is restricted to a one-mode distribution?

$$q^* = \arg\min_q D_{\mathrm{KL}}(p \| q)$$

Asymmetric KL Divergence

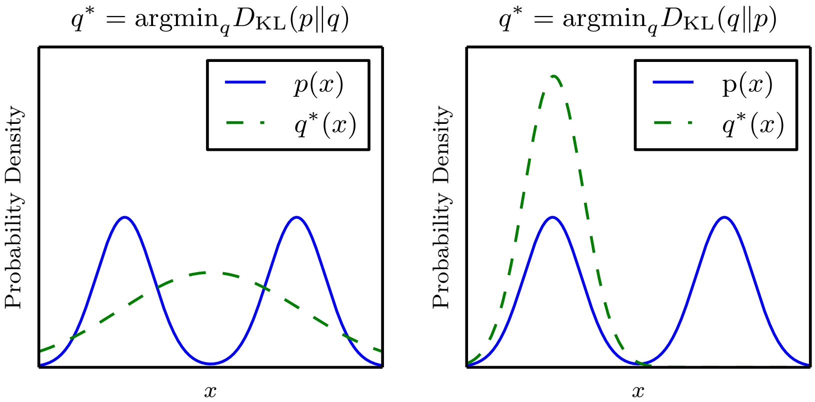


Figure: Left: emphasize on ”average mode“; Right: emphasize on the most probable mode.

Asymmetric KL Divergence

For the left plot:

- $q^* = \arg\min_q \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$.
- Wherever $p(x) > 0$, $q(x)$ has to > 0, making $q(x)$ cover all modes of $p(x)$.
- This is called over-estimating of $p(x)$, which is the case in expectation propagation (EP).

For the right plot:

- $q^* = \arg\min_q \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x$.
- Wherever $p(x) = 0$, $q(x)$ has to $= 0$, making $q(x)$ cover only one mode of $p(x)$.
- This is called under-estimating of $p(x)$, which is the case in variational inference (VI).