Probability

CMPT 498/898 Deep Learning and Applications

Najeeb Khan
najeeb.khan@usask.ca

Slides based on Deep Learning, Goodfellow et al. 2016 (Ch. 3)

Winter 2020

Introduction

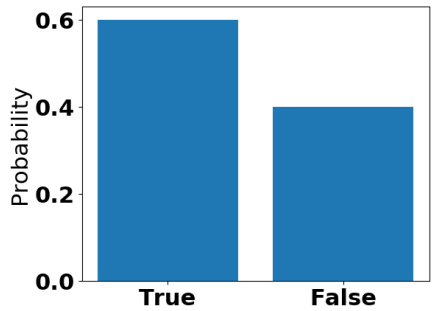
- Machine learning must always deal with uncertain quantities
- Possible sources of uncertainty:
  - Inherent stochasticity (Quantum mechanics, perfectly shuffled cards)
  - Incomplete observability
  - Incomplete modeling (Discretized object localization)
- Probability provides a means of quantifying uncertainty
- Helps determine the likelihood of a proposition being true given the likelihood of other propositions

Two Interpretations

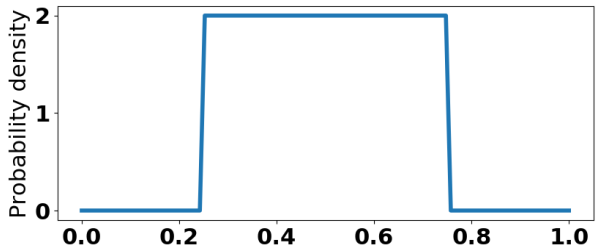
- Frequentist Probability
  - Analyze the frequencies of events (Drawing a certain hand of cards in a game of poker)
  - The analyzed events are often repeatable
- Bayesian Probability
  - Use probability to represent a degree of belief (what are the chances of a patient having a flu)
  - Applicable to propositions that are not repeatable

Random Variable

- Random variable is a variable that can take on different values randomly
- Probability distribution: a description of how likely a random variable is to take on each of its possible states
  - Discrete states: probability mass function



  - Continuous states: probability density function



Probability Distributions

- Probability Mass function P:
  1. Domain of P must be the set of all possible states of the random variable x
  2. $$\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1.$$
  3. $$\sum_{x \in \mathrm{x}} P(x) = 1.$$
- Probability Density function p:
  1. Domain of p must be the set of all possible states of the random variable x
  2. $$\forall x \in \mathrm{x}, p(x) \geq 0$$
  3. $$\int_{x \in \mathrm{x}} p(x) dx = 1.$$

Marginal Probability

- Joint probability mass function $P(\mathrm{x}, \mathrm{y})$

$$P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y)$$

- Joint PDF $p(\mathrm{x}, \mathrm{y})$

$$p(x) = \int p(x, y) dy$$

Chain rule of Probabilities

$$P(a, b) = P(a)P(b|a),$$

where $P(b|a)$ is the conditional probability of b given a has happened.

$$P(a, b, c) = P(a)P(b|a)P(c|a, b)$$

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

Bayes' Rule

$$P(a, b) = P(a)P(b|a) = P(b)P(a|b)$$

$$P(a|b) = \frac{P(a)P(b|a)}{P(b)}$$

$$P(a|b) = \frac{P(a)P(b|a)}{\sum_a P(b|a)P(a)}$$

Independence

- Independent random variables

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

- Conditionally independent variables

$$P(x, y|z) = p(x|z)p(y|z)$$

Expectation

- A measure of the average value of the random variable
- Discrete random variables

$$\mathbb{E}_{\mathrm{x}\sim P}\big[f(x)\big] = \sum_{x} P(x)f(x)$$

- Continuous random variables

$$\mathbb{E}_{\mathrm{x}\sim P}\big[f(x)\big] = \int P(x)f(x)dx$$

- Linearity

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

Variance

- A measure of the spread of samples of a random variable

$$\mathrm{Var}(f(x)) = \mathbb{E}\Big[\big(f(x)\big)^2\Big] \quad \text{when } \mathbb{E}[f(x)] = 0$$

- Non-zero mean

$$\mathrm{Var}(f(x)) = \mathbb{E}\Big[\big(f(x) - \mathbb{E}[f(x)]\big)^2\Big]$$

Covariance

A measure of linear relationship between two random variables

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\big[f(x)g(y)\big]$$

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\Big[\big(f(x) - \mu_{g(x)}\big)\big(g(y) - \mu_{g(y)}\big)\Big]$$

$$\mu_x = \mathbb{E}[x]$$

Covariance matrix of $\mathbf{x} \in \mathbb{R}^n$

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

Correlation: normalized covariance

Common Probability Distributions

Bernoulli: distribution over a single binary random variable

$$P(\mathrm{x} = x) = \phi^x(1-\phi)^{1-x}$$

Expectation: $\mathbb{E}[x] = \phi$
Variance: $\mathrm{Var}(x) = \phi(1-\phi)$

Gaussian: most commonly used distribution over real numbers

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big)$$

where $\mathbf{x} \in \mathbb{R}^n$
$\boldsymbol{\mu}$ is the mean vector
$\boldsymbol{\Sigma}$ is the covariance matrix

Mixture Distribution

$$P(x) = \sum_i P(c = i)P(x|c = i)$$

On each trial, choose c according to P_c and sample from the corresponding component distribution $P(x|c)$

Information Theory

Goal: Quantifying how much information is present in a signal

- Less likely events have higher information content
- Independent events have additive information

Self-information of sample x:

$$I(x) = \log \frac{1}{P(x)} = -\log P(x)$$

Nat (unit): information gained by observing an event of probability $1/e$

Entropy of a probability distribution P:

$$H(\mathrm{x}) = -\mathbb{E}_{x \sim P}\big[\log P(x)\big]$$

H is the expected amount of information in an event drawn from P

Information Theory

Entropy of Bernoulli distribution as a function of ϕ

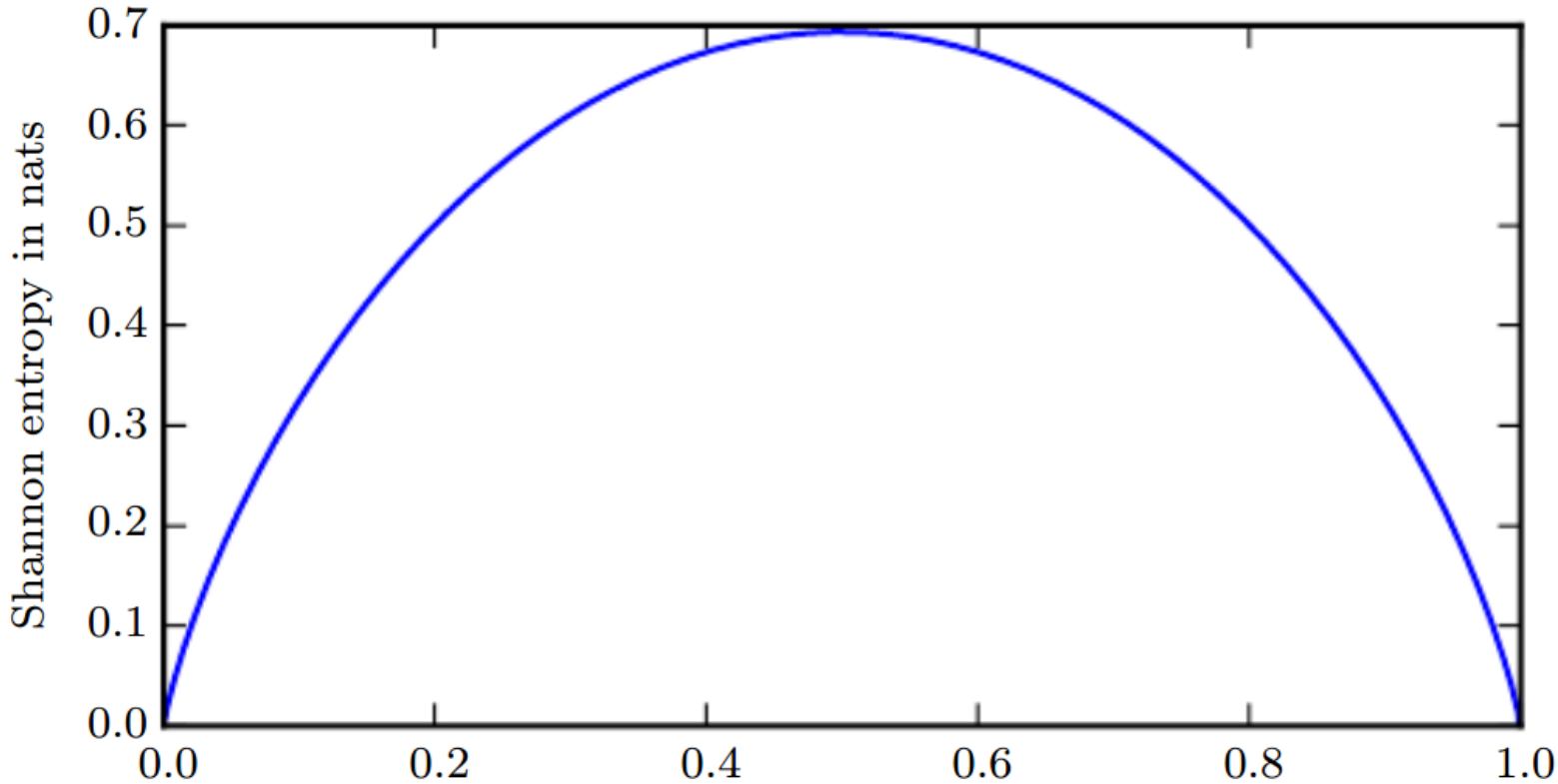


Image Source: Deep Learning, Goodfellow et al. Fig 3.5

Information Theory

Relative Entropy or Kullback–Leibler (KL) divergence

Given two distributions P and Q over x

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

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x\in\mathcal{X}} P(x)\log P(x) - \underbrace{\sum_{x\in\mathcal{X}} P(x)\log Q(x)}_{\text{Cross entropy H(P,Q)}}$$

$$D_{\mathrm{KL}}(P \parallel Q) \geq 0$$

$$D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$$