Probability and Inference

Artificial Intelligence

Dongsuk Yook
Artificial Intelligence Laboratory
Korea University

Contents

- Review of Probabilities and Statistics
- Probabilistic Inference

Class Objectives

- Understanding the fundamentals of probability theory
- Being able to implement naive Bayes classifiers

Contents

Introduction

- Random experiment
  - Nondeterministic outcomes
  - e.g., coin toss, die roll

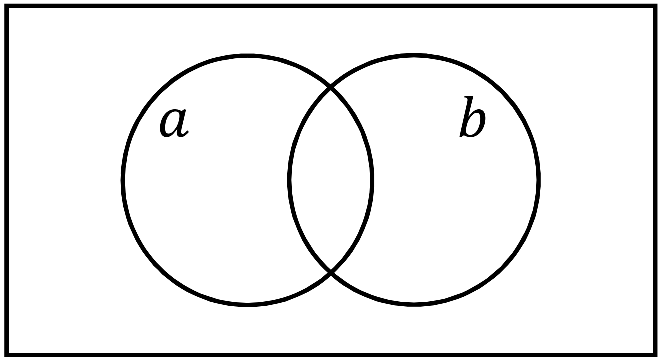
- Sample space: Ω
  - A set of all possible outcomes
  - Mutually exclusive and exhaustive
  - e.g., die roll: 1, 2, …, 6

- Event
  - A subset of the sample space
  - e.g., even numbers in rolling a die, doubles in rolling two dice

- Probability
  - $P(e)$: the probability of an event e
  - Proportion (or relative frequency)
  - Degree of belief

Axioms of Probability

- Kolmogorov's axioms
  - $0 \le P(\omega) \le 1$ for every $\omega \in \Omega$
  - $\sum_{\omega \in \Omega} P(\omega) = 1$
  - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$; inclusion-exclusion principle



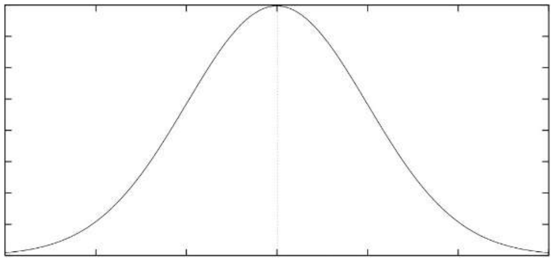
Random Variables and Probability Mass/Density

- Random variable
  - A function that maps from Ω (*domain*) to a set of possible values (*range*)
  - Discrete
  - Continuous

- Probability mass function (pmf)
  - Probability of a discrete random variable
  - e.g., $\underbrace{P(X=0)}_{P(0)} = 0.4,\ \underbrace{P(X=1)}_{P(1)} = 0.6$; X: random variable

- Probability distribution
  - e.g., $\mathbf{P}(X) = \langle 0.4, 0.6 \rangle$; Bernoulli distribution
    categorical distribution

- Probability density function (pdf)
  - Probability *density* of a continuous random variable
  - $\underbrace{P(X=x)}_{P(x)} = \lim_{dx \to 0} P(x \le X \le x + dx)/dx$

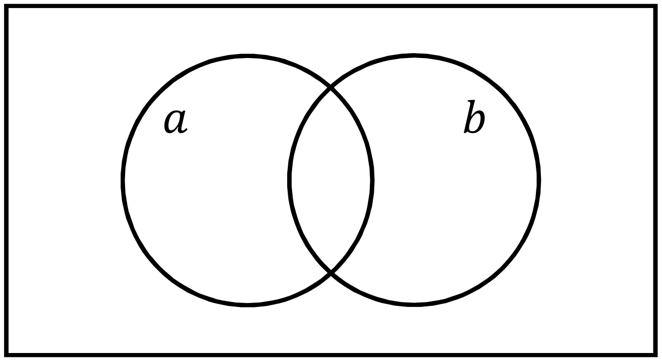


- Cumulative distribution function (also called *probability distribution function*: PDF)
  - $F(x) = P(X \le x) = \sum_{u \le x} P(u)$; $\int_{-\infty}^{x} P(u)du$

Prior/Posterior/Joint/Marginal Probabilities

- Prior probability (also called *unconditional probability* or just *prior* for short)
  - $\underbrace{P(X = a)}_{P(a)}$

- Posterior probability (also called *conditional probability* or just *posterior* for short)
  - $\underbrace{P(X = a|Y = b)}_{P(a|b)} \equiv \frac{P(X=a,Y=b)}{P(Y=b)} \Big\} \frac{P(a,b)}{P(b)}$



- Joint probability
  - $\underbrace{P(X = a, Y = b)}_{P(a,b)} \equiv \underbrace{P(X = a \wedge Y = b)}_{P(a \wedge b) \text{ or } P(a \cap b)}$

- Product rule
  - $P(a, b) = P(a|b)P(b) = P(b|a)P(a)$

- Marginal probability
  - $\underbrace{P(X = a)}_{P(a)} = \sum_{y \in Y} \underbrace{P(X = a, Y = y)}_{P(a,y)}$; marginalization, summing out
  - $\underbrace{P(X = a)}_{P(a)} = \sum_{y \in Y} \underbrace{P(X = a|Y = y)P(Y = y)}_{P(a|y)\, P(y)}$; conditioning

Chain Rule

- Chain rule
  - $$\begin{aligned} & P(x_1, x_2, \cdots, x_{n-1}, x_n) \\ & = P(x_n | x_1, \cdots, x_{n-1}) P(x_1, \cdots, x_{n-1}) \\ & = P(x_n | x_1, \cdots, x_{n-1}) P(x_{n-1} | x_1, \cdots, x_{n-2}) P(x_1, \cdots, x_{n-2}) \\ & \vdots \\ & = P(x_n | x_1, \cdots, x_{n-1}) P(x_{n-1} | x_1, \cdots, x_{n-2}) \cdots P(x_2 | x_1) P(x_1) \\ & = \prod_{i=1}^{n} P(x_i | x_1, \cdots, x_{i-1}) \qquad ; P(x_1 | x_0) \equiv P(x_1) \end{aligned}$$

$P(a \wedge b) = P(a|b)P(b)$

Independence

- Independence (also called *absolute independence* or *marginal independence*)
  - $P(x|y) = P(x)$
  - $P(x,y) = P(x|y)P(y) = P(x)P(y)$

- Conditional independence
  - $P(x|y) \neq P(x)$
  - $P(x|z) \neq P(x)$
  - $P(x|y,z) = P(x|z)$
  - $P(x,y|z) = P(x|y,z)P(y|z) = P(x|z)P(y|z)$; $P(x,y) \neq P(x)P(y)$

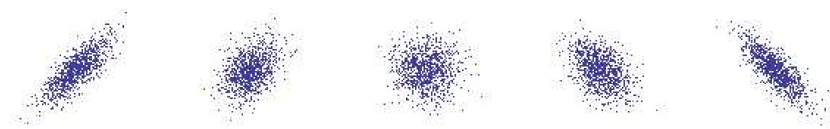
$P(a \wedge b) = P(a|b)P(b)$

Expectation

- Expectation (also called *mean*) of a random variable X
  - $\mathbb{E}(X) \equiv \sum_x xP(x)$; $\mu \equiv E(X)$
  - $\mathbb{E}(X) \equiv \int_{-\infty}^{\infty} xP(x)\, dx$

- Variance: σ^2
  - $\mathrm{VAR}(X) \equiv \mathbb{E}\left((X-\mu)^2\right)$; $\sigma^2 \equiv \mathrm{VAR}(X)$

- Covariance of two random variables X and Y
  - $\mathrm{COV}(X,Y) \equiv \mathbb{E}\left((X-\mu_X)(Y-\mu_Y)\right)$

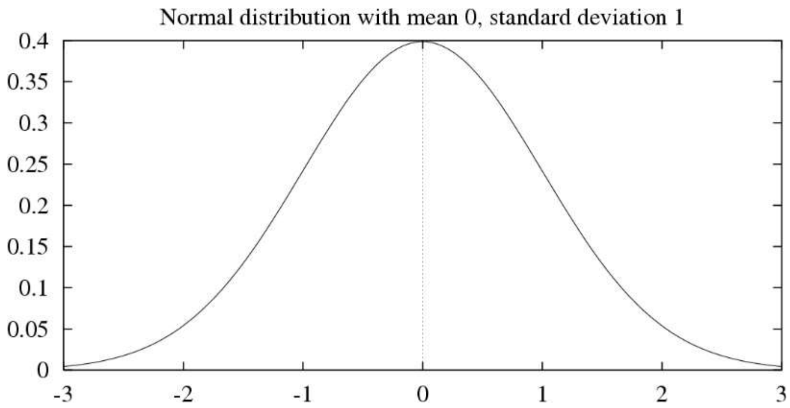


- Covariance matrix, $\mathbf{\Sigma}$, a random vector $\mathbf{X}$
  - $\Sigma_{ij} \equiv \mathrm{COV}(X_i, X_j) = \mathbb{E}\left((X_i-\mu_i)(X_j-\mu_j)\right)$

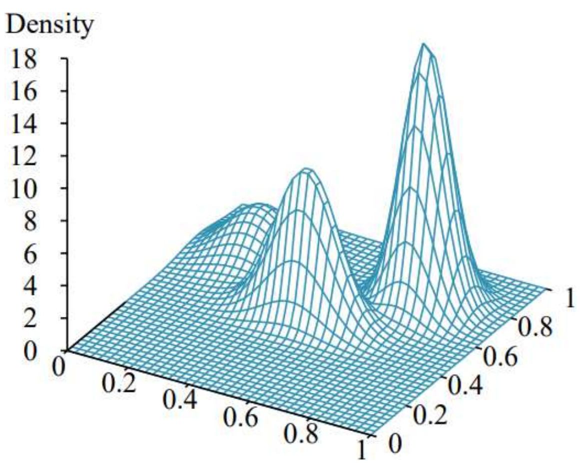
- Expectation of a random vector $\mathbf{X}$
  - $\mathbb{E}(\mathbf{X}) \equiv \sum_{\mathbf{x}} \mathbf{x}P(\mathbf{x})$
  - $\mathbb{E}(\mathbf{X}) \equiv \int_{-\infty}^{\infty} \mathbf{x}P(\mathbf{x})\, d\mathbf{x}$

Gaussian Distribution

- Gaussian distribution (also called *normal distribution*)
  - $P(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



- Multivariate Gaussian distribution (also called *multivariate normal distribution*)
  - $P(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$



Central Limit Theorem

- The distribution formed by sampling n independent random variables and taking their mean tends to a normal distribution as n tends to infinity.

Contents

- Review of Probabilities and Statistics
- Probabilistic Inference

Probabilistic Inference

- Probabilistic inference
  - The computation of posterior probabilities for query propositions given observed evidence
  - $P(x|\mathbf{e})$; $\mathbf{e} \equiv e_1, e_2, \cdots, e_n$
    $= P(x, \mathbf{e})/P(\mathbf{e})$
    $= \alpha P(x, \mathbf{e})$; $\alpha \equiv 1/P(\mathbf{e})$
    $= \alpha \sum_{\mathbf{u}} P(x, \mathbf{e}, \mathbf{u})$; $\mathbf{u}$: unobserved

- Bayes' rule (also called *Bayes' law* or *Bayes' theorem*)
  - $$\underbrace{P(c|e)}_{\text{posterior}} = \frac{\overbrace{P(e|c)}^{\text{likelihood}}\ \overbrace{P(c)}^{\text{prior}}}{\underbrace{P(e)}_{\text{evidence}}}$$
  - $$P(c|e,u) = \frac{P(e|c,u)P(c|u)}{P(e|u)}$$

$P(a \wedge b) = P(a|b)P(b)$
$P(a \wedge b) = P(b|a)P(a)$
$P(a|b) = P(b|a)P(a)/P(b)$

- $$\underbrace{P(\text{cause}|\text{effect})}_{\text{diagnostic direction}} = \frac{\overbrace{P(\text{effect}|\text{cause})}^{\text{causal direction}}P(\text{cause})}{P(\text{effect})}$$

Naive Bayes Models

- Naive Bayes model (also called *naive Bayes classifier*)
  - $P(c|e_1, e_2, \cdots, e_n)$
    $= P(e_1, e_2, \cdots, e_n|c)P(c)/P(e_1, e_2, \cdots, e_n)$; Bayes' theorem
    $= \alpha P(e_1, e_2, \cdots, e_n|c)P(c)$; $\alpha \equiv 1/P(e_1, e_2, \cdots, e_n)$
    $= \alpha P(c) \prod_i P(e_i|c)$; conditional independence
  - $P(c|e_1, e_2, \cdots, e_n)$
    $= P(c, e_1, e_2, \cdots, e_n)/P(e_1, e_2, \cdots, e_n)$
    $= \alpha P(c, e_1, e_2, \cdots, e_n)$; $\alpha \equiv 1/P(e_1, e_2, \cdots, e_n)$
    $= \alpha \sum_{u_1,\cdots,u_m} P(c, e_1, e_2, \cdots, e_n, u_1, \cdots, u_m)$; $u_1, \cdots, u_m$: unobserved
    $= \alpha \sum_{u_1,\cdots,u_m} P(c)P(e_1, e_2, \cdots, e_n, u_1, \cdots, u_m|c)$; product rule
    $= \alpha \sum_{u_1,\cdots,u_m} P(c)[\prod_i P(e_i|c)]P(u_1, \cdots, u_m|c)$; conditional independence
    $= \alpha P(c) \prod_i P(e_i|c) \sum_{u_1,\cdots,u_m} P(u_1, \cdots, u_m|c)$
    $= \alpha P(c) \prod_i P(e_i|c)$

$$P(c|e) = \frac{P(e|c)P(c)}{P(e)}$$

$$P(c|e,u) = \frac{P(e|c,u)P(c|u)}{P(e|u)}$$

Example: Text Classification

- Text classification with a naive Bayes model
  - $\mathbf{P}(Class|word_1, word_2, \cdots, word_n) = \alpha \mathbf{P}(Class) \prod_i \mathbf{P}(word_i|Class)$
  - $\arg\max_{class} P(class|word_1, word_2, \cdots, word_n)$
    $= \arg\max_{class} \frac{P(word_1, word_2, \cdots, word_n|class) P(class)}{P(word_1, word_2, \cdots, word_n)}$
    $= \arg\max_{class} P(class) \prod_i P(word_i|class)$
  - Bag-of-words: $word_1, word_2, \cdots, word_n$
  - e.g., news, sports, business, weather, entertainment
    - *<u>Stocks</u> rallied on Monday, with major indexes gaining 1% as optimism persisted over the first <u>quarter</u> <u>earnings</u> season.*
    - *Heavy <u>rain</u> continued to pound much of the east <u>coast</u> on Monday, with <u>flood</u> warnings issued in New York City and other locations.*

$P(c|e_1, e_2, \cdots, e_n)$
$= P(e_1, e_2, \cdots, e_n|c) P(c) / P(e_1, e_2, \cdots, e_n)$
$= \alpha P(c) \prod_i P(e_i|c)$

Summary and Preview

- Review of Probabilities and Statistics
  - Introduction
  - Axioms of Probability
  - Random Variables and Probability Density
  - Prior, Posterior, Joint, Marginal Probabilities
  - Chain Rule
  - Independence
  - Expectation
  - Gaussian Distribution
  - Central Limit Theorem
- Probabilistic Inference
  - $P(x|\mathbf{e}) = \alpha \sum_{\mathbf{u}} P(x, \mathbf{e}, \mathbf{u})$
  - $P(c|e_1, e_2, \cdots, e_n) = \alpha P(c) \prod_i P(e_i|c)$
- Bayesian Networks