Math Background (I)

杜俊
jundu@ustc.edu.cn

Review

- Many applications fall into pattern classification problem.
- Pattern classification: based on some observed information *X* of an input *P*, classify it into one of the finite number of classes C_i.
  - Automatic speech recognition
  - Handwriting recognition
  - Optical character recognition
- Paradigm shift for pattern classification
  - Rule/knowledge based method
  - Data-driven method
- Statistical pattern classification
  - Feature extraction
  - Statistical modeling and decision making

Math Review

- Probability & Statistics
  - Bayes' theorem
  - Random variables: discrete vs. continuous
  - Probability distribution: PDF and CDF
  - Statistics: mean, variance, moment
  - Parameter estimation: MLE
- Information Theory
  - Entropy, mutual information, information channel, KL divergence
- Function Optimization
  - Constrained/unconstrained optimization
- Linear Algebra
  - Matrix manipulation

Probability Definition

- Sample space Ω
  - Collection of all possible observed outcomes
- Event $A: A \subseteq \Omega$ including null event ϕ
- σ-field F_Ω : set of all possible events $A \in F_\Omega$
- Probability function (measurable) $P: F_\Omega \to [0,1]$
  - $P(\phi) = 0 \qquad P(\Omega) = 1$
  - $A \subseteq B \Rightarrow P(A) \le P(B)$
  - $A \cap B = \phi \Rightarrow P(A \cup B) = P(A) + P(B)$

Examples

- Example I: experiment to toss a 6-face dice once
  - Sample space: {1,2,3,4,5,6}
  - Events: X={even number}, Y={odd number}, Z={larger than 3}.
  - σ-field: set of all possible events
  - Probability function: relative frequency

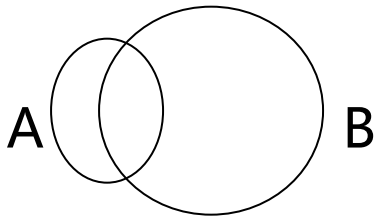
- Example II
  - Sample space: {x: x is the height of a person on earth}
  - Events: A={x: x>200cm} B={x: 120cm<x<130cm}
  - σ-field: set of all possible events
  - Probability function

$$\Pr(A) = \frac{\#\text{ of persons whose height over 200cm}}{\text{total }\#\text{ of persons in the earth}}$$

Conditional Events

- Prior/Marginal Probability
  - Probability of an event before considering any additional knowledge or observing any other events (or samples): $P(A)$
- Joint Probability
  - Probability of several events occurring concurrently: $P(A \cap B)$
- Conditional Probability
  - Probability of one event after another event has occurred $P(A|B)$
  - $P(A|B) = P(A \cap B) / P(B)$
- Two Rules:
  - Addition rule:

    $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



  - Multiplication rule:

    $P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2 \mid A_1) \cdots P(A_n \mid \bigcap_{i=1}^{n-1} A_i)$

Bayes' Theorem

- Swapping dependency between events

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A)}$$

- In some cases, not important to compute $P(A)$

$$B^* = \arg\max_B P(B \mid A) = \arg\max_B \frac{P(A \mid B)P(B)}{P(A)} = \arg\max_B P(A \mid B)P(B)$$

- Another form of Bayes' theorem

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} = \frac{P(A \mid B)P(B)}{\sum_{B'} P(A \mid B')P(B')}$$

- The theoretical basis of pattern classification problem

Random Variable

- A random variable (R.V.) is a variable which could take various values with different probabilities. A random vector is a vector whose elements are all random variables

- PMF(Probability Mass Function) for discrete R.V.

$$f(x) = \Pr(X = x) \quad \text{for } x = x_1, x_2, \cdots \qquad \sum_{x_i} f(x_i) = 1$$

- CDF and PDF for continuous R.V.
  - CDF(Cumulative Distribution Function)

$$F(x) = \Pr(X \le x) \quad (-\infty < x < \infty) \quad \lim_{x \to -\infty} F(x) = 0 \quad \lim_{x \to \infty} F(x) = 1$$

  - PDF(Probability Density Function)

$$\forall a, b \quad \Pr(a \le X \le b) = \int_a^b f(x)\,\mathrm{d}x \quad F(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x \quad \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$$

Statistics

- Expectations of R.V. and its functions

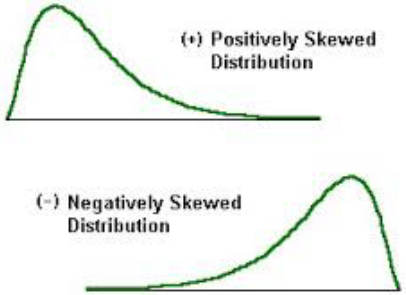
$$E(X)=\int_{-\infty}^{\infty} x\cdot f(x)\mathrm{d}x \qquad \text{or} \qquad \sum_i x_i\cdot p(x_i)$$

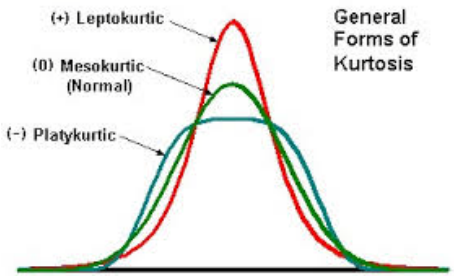
$$E(q(X))=\int_{-\infty}^{\infty} q(x)\cdot f(x)dx \quad \text{or} \quad \sum_i q(x_i)\cdot p(x_i)$$

- Mean and variance

$$\mathrm{Mean}(X)=E(X) \quad \mathrm{Var}(X)=E([X-E(X)]^2)$$

- Raw moment $E(X^r)$ and central moment $E([X-E(X)]^r)$
  - Skewness（偏度）and Kurtosis（峰度）





Applications

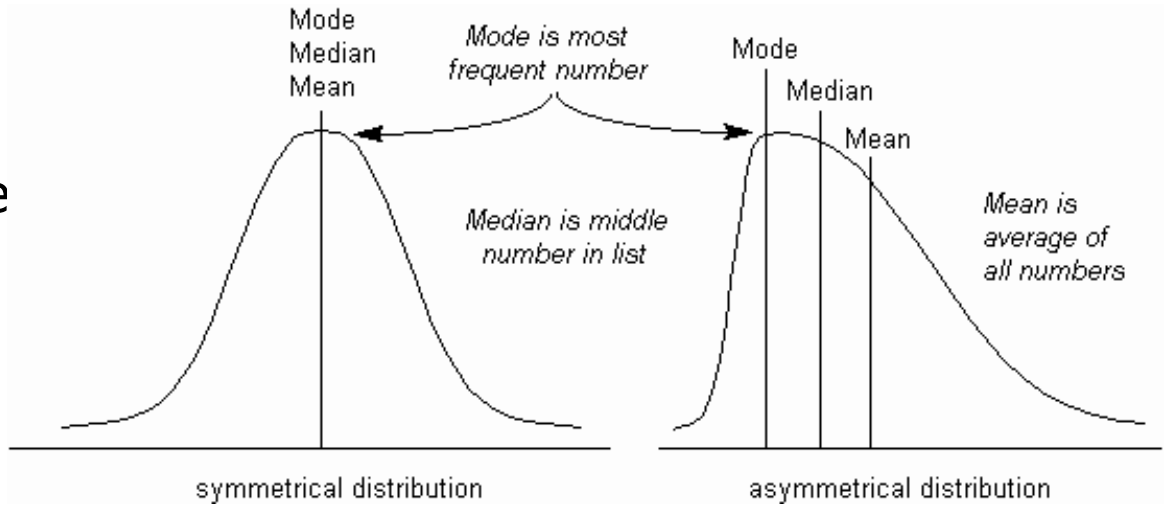
- Feature Transformation
  - CDF Matching: Histogram equalization

$$\Pr(X \le x) = \Pr(Y \le y(x))$$

  - Moment Matching

$$M(x) = M(f(y))$$

- Point Estimation
  - Mean, Median, Mode



Joint/Marginal/Conditional Distribution

- Joint CDF/PDF of two continuous random variables

$$p(x,y) = \Pr(X \le x, Y \le y)$$

$$\Pr(a \le x \le b, c \le y \le d) = \int_a^b \int_c^d f(x,y)\,dy\,dx$$

- Marginal PDF $f(x) = \int f(x,y)dy$
- Conditional PDF $f(x \mid y) = f(x,y) / f(y)$

Conditional Distribution

- Conditional Expectation

$$E(q(X)\,|\,Y=y_0)=\int_{-\infty}^{\infty}q(x)\,f(x\,|\,y_0)dx$$

- Conditional Mean（MMSE Estimator）

$$E(X\,|\,Y=y_0)=\int x\cdot f(x\,|\,y_0)dx$$

- Covariance of between two random variables

$$\mathrm{Cov}(X,Y)=E([X-E(X)][Y-E(Y)])=\int_x\int_y(x-E(X))(y-E(Y))\cdot f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

- Independence

$$f(x,y)=f(x)f(y)\quad f(x\,|\,y)=f(x)$$

- Uncorrelation

$$\mathrm{Cov}(X,Y)=0$$

Useful Distributions

- Discrete Distributions
  - Binomial distribution
  - Multinomial distribution (多项分布)
  - Multinomial mixture distribution

- Continuous Distributions
  - Uniform distribution
  - Poisson distribution
  - Beta distribution
  - Dirichlet distribution
  - Gamma distriubution
  - Univariate/Multivariate Gaussian distribution
  - Univariate/Multivariate Gaussian mixture distribution

Parametric/Non-Parametric Distribution

- Parametric Distribution
  - A R.V. described by a small number of parameters in PDF/PMF
  - E.g., Multinomial distribution, Gaussian distribution
  - Many useful and known parametric distributions
  - Parameter Estimation: Maximum likelihood estimation, etc.

- Non-Parametric Distribution
  - Usually described by the data samples themselves
  - Sampling distribution and histogram
  - Kernel density estimation
  - Nearest neighbor density estimation

Function of Random Variables

- Function of R.V.s is also a R.V.
- Problem easier for known and popular R.V.s
  - The sum of several independent Gaussian R.V.s
- Law of large numbers
- Central-limit theorem
- Transformation of R.V.s

$$\vec{X} = (X_1, \cdots X_n) \text{ and } \vec{Y} = (Y_1, \cdots, Y_n)$$

$$Y_1 = g_1(\vec{X}), \cdots, Y_n = g_n(\vec{X})$$

$$X_1 = h_1(\vec{Y}), \cdots, X_n = h_n(\vec{Y})$$

$$J(\vec{Y}) = \begin{bmatrix} \frac{\partial h_1}{\partial Y_1} & \cdots & \frac{\partial h_1}{\partial Y_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial h_n}{\partial Y_1} & \cdots & \frac{\partial h_n}{\partial Y_n} \end{bmatrix}$$

$$\Rightarrow p_Y(\vec{Y}) = p_X(h_1(\vec{Y}), \cdots h_n(\vec{Y})) \cdot \left| J(\vec{Y}) \right|$$

Maximum Likelihood Estimation (I)

- Principle: we view the parameters as quantities whose values are fixed but unknown. The best estimate of their value is defined to be the one that maximizes the probability of observing the samples actually observed.
  - Best interpret the data
  - Fit the data best
- The likelihood function
  - $p(X \mid \theta)$ → data distribution PDF of different X if θ is given
  - $p(X \mid \theta)$ → likelihood function of θ if data X is given
- Assume D contains n samples, $X_1, X_2, \ldots, X_n$, since the samples were drawn independently from $p(X \mid \theta)$. Then maximum-likelihood estimate of θ is the value θ_{ML} that maximizes $p(D|\theta)$

$$\theta_{ML} = \arg\max_{\theta} p(D \mid \theta) = \arg\max_{\theta} \prod_{k=1}^{n} p(X_k \mid \theta)$$

Maximum Likelihood Estimation (II)

- Equivalent to maximize the log-likelihood function

$$\theta_{ML} = \arg\max_{\theta} l(\theta) = \arg\max_{\theta} \sum_{k=1}^{n} \ln p(X_k \mid \theta)$$

- Examples
  - Gaussian distribution
    - Closed-form solution for mean and variance parameters
  - Gaussian mixture distribution (model)
    - No closed-form solution
    - EM (Expectation-Maximization) algorithm

Probability Theory Recap

- Probability Theory Tools
  - Fuzzy description of phenomena
  - Statistical modeling of data for inference
- Statistical Inference Problems
  - Classification: choose one of the stochastic sources
  - Regression: fit a model to a given set of samples
- Parametric vs. Non-Parametric Distributions
  - Parsimonious or extensive description (model vs. data)
  - Sampling, data storage and sufficient statistics
- Real-World Data vs. Ideal Distributions
  - There is no perfect goodness-of-fit
  - Ideal distributions are used for approximation
  - Sum of random variables and law of large numbers