Review of Mathematical Foundations

Calculus, Set Theory, and Linear Algebra

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Review basic notations from Calculus & Set Theory

Objective

Review key Linear Algebra concepts and operations

Basic Notations from Calculus (1/3)

- **Derivative of *f*(*x*) with respect to *x***

- **Partial derivative of a function *f*(*x*,*y*,…) with respect to *x***

  – Note: the function may be scalar-valued or vector-valued

Basic Notations from Calculus (2/3)

- $\Re^{d}$: *d*-dimensional Euclidean space.

- Gradient operator in $\Re^{\mathrm{d}}$: ∇

Basic Notations from Calculus (3/3)

- The integral of *f*(*x*) between *a* and *b*

- The argmin or argmax notation

Basic Notations from Set Theory (1/2)

A set *S* is a collection of objects.

- ∅: the empty set (a special set that contains no object)

Some basic relations and operations

- $x \in A$: An object *x* is a member of a set *A*.
- $A \subseteq B$: Set *A* is a *subset* of *B* $\Leftrightarrow x \in A \Rightarrow x \in B$
- $B \subset C$: Set *B* is a *proper subset* of C.

Basic Notations from Set Theory (2/2)

Some basic relations and operations

- $A \cup B$: The union of *A* and *B.*
- $A \cap B$: The intersection of *A* and *B.* (*AB* in shorthand)
- A^c or $\overline{A}$: The complement of *A*
- *A* and *B* are disjoint if $A \cap B = \emptyset$

Linear Algebra: Basic Notations (1/4)

- A *d*-dimensional column vector x and its transpose x^t

- *n* by *d* matrix M and its *d* by *n* transpose M^t

Linear Algebra: Basic Notations (2/4)

- A square matrix M is symmetric if
- Multiplying a vector by a matrix: Mx = y
- Multiplying two matrices M_1 and M_2

Linear Algebra: Basic Notations (3/4)

- The identity matrix I of *d* by *d*
- Inner product of two vectors $x^t y$
- Outer product of two vectors $x y^t$

Linear Algebra: Basic Notations (4/4)

- The length or Euclidean norm of a vector x, denoted ||x||

- Normalized vector, ||x|| = 1

Matrix: Additional Definitions (1/2)

- **Determinant of a matrix M: denoted |M| or det(M)**
  - Look at size 2x2
  - What about size 3x3 and above?
- **Trace of a matrix**

Matrix: Additional Definitions (2/2)

- Matrix inversion M^{-1}
- Eigenvectors and eigenvalues of M

Derivatives Involving Matrices (1/3)

- If the entries of a matrix M depend on a scalar parameter θ, we have $\frac{\partial \mathbf{M}}{\partial \theta} =$

- Derivative of a scalar-valued function $f(\mathrm{x})$ of d variables x_i, i=1,…,d, and x=(x_1, …, x_d)^t , or the gradient w.r.t. x is $\nabla f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} =$

Derivatives Involving Matrices (2/3)

- **If f(x) is an *n*-dimensional vector-valued function of *d* variables x_i, *i*=1,…,*d*, and x=(x_1, …, x_d)^t , we have the derivative as*** $\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} =$

* We could use the Jacobian form too; See “numerator layout” vs “denominator layout” in matrix calculus.

Derivatives Involving Matrices (3/3)

Some useful results:

$$\frac{\partial}{\partial \mathbf{x}}[\mathrm{M}\mathbf{x}] =$$

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{y}^{\mathrm{t}}\mathbf{x}] =$$

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^{t}\mathrm{M}\mathbf{x}] =$$

Review of Mathematical Foundations

Basics in Probability Theory

Ira A. Fulton Schools of Engineering

Arizona State University

Objective

Objective

Define Probability Space

Objective

Discuss Conditional Probability and Bayes Rule

Probability Space (1/2)

A probability space is a triplet (Ω, $\mathcal{B}$, P) that is used to model a process or an experiment with random outcomes.

- The **sample space** Ω is the set of all possible outcomes of an experiment
  - Consider two different experiments

    (1) Tossing a coin; (2) Tossing a die

Probability Space (2/2)

- **$\mathcal{B}$: a sigma algebra (or Borel field), or informally, a collection of subsets of Ω, subject to some constraints (like containing the empty set, being closed under complements and countable union)**

- ***P:* a measure called probability defined on $\mathcal{B}$, that satisfies**
  - $P(A) \geq 0$ for all $A \in \mathcal{B}$
  - $P(\Omega) = 1$
  - If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint then $P(\cup A_i) = \sum P(A_i)$ (i.e., $A_j A_k = \varnothing, \forall j \neq k$)

Conditional Probability

- Let $(\Omega, \mathcal{B}, P)$ be a probability space, and let $H \in \mathcal{B}$ with $P(H)>0$. For any $B \in \mathcal{B}$, we define $P(B|H) = P(BH) / P(H)$ and call $P(B|H)$ the conditional probability of B, given H.

The Total Probability Rule

- Let $(\Omega, \mathcal{B}, P)$ be a probability space, and let $\{H_j\}$ be pairwise disjoint events in $\mathcal{B}$ (i.e., $H_jH_k = \varnothing$, $\forall j \neq k$) and $\cup_{j=1,\ldots,\infty} H_j = \Omega$. Suppose $P(H_j) > 0$, $\forall j$, then $P(B) = \sum_{j=1,\ldots,\infty} P(H_j)P(B|H_j)$
  - Such $\{H_j\}$ is called a partition of Ω.

The Bayes Rule

- **Let ($\mathcal{\Omega}$, $\mathcal{B}$, *P*) be a probability space, and let {H_j} be pairwise disjoint events in $\mathcal{B}$ with $\cup_{j=1,\ldots,\infty} H_j = \Omega$, and $P(H_j)>0$, $\forall j$. We have, $\forall B \in \mathcal{B}$ and $P(B)>0$,**

$$P(H_j|B) = \frac{P(H_j)\, P(B\,|H_j)}{\sum_{i=1,\ldots,\infty} P(H_i)P(B|H_i)}, \qquad \forall\, j$$

Independence of Events

- Let $(\Omega, \mathcal{B}, P)$ be a probability space, $\forall A, B \in \mathcal{B}$, we say A and B are independent if $P(AB) = P(A)P(B)$.

Review of Mathematical Foundations

Random Variables and Common Distributions

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Review random variables & their distributions

Discrete Random Variables

- **Let x be a discrete random variable that can take any of the m different values in the set V={v_1, v_2, …, v_m} with respective probabilities {p_1, p_2, …, p_m}, i.e., p_i=$Prob[x= v_i]$.**
  - $p_i \geq 0$, $\sum_{j=1,\ldots,m} p_j = 1$

- **Probability Mass Function $P(x)$ is used to represent the set of probabilities {p_1, p_2, …, p_m}**
  - $P(x) \geq 0$, $\sum_{x \text{ in } V} P(x) = 1$

Expected Value (Means) & Variance

- The expected value (mean) of *x*, E[*x*], often denoted μ

$$\mu = E[x] = \sum_{x \text{ in } V} xP(x)$$

- The expected value of a function *f*(*x*), E[*f*(*x*)],

$$E[f(\mathrm{x})] = \sum_{x \text{ in } V} f(x)P(x)$$

- E[ ] is linear when viewed as an operator.

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

- The variance of *x*, Var[*x*], often denoted σ^2

$$\sigma^2 = \mathrm{Var}(x) = E[(x-\mu)^2] = \sum_{x \text{ in } V} (x-\mu)^2 P(x)$$

Joint Distributions

- **Consider a pair of discrete random variables, *x* and *y*, taking values in V={v$_1$, v$_2$, …, v$_m$} and W={w$_1$, w$_2$, …, w$_n$} respectively.**
  - (x, y) to take a pair of values (v_i, w_j) with probability p_{ij}
  - Or, we consider the **joint probability mass function** $P(x, y)$

Marginal Distributions

Knowing $P(x, y)$, can we figure out $P_x(x)$ or $P_y(y)$?

➔ The concept of **marginal distribution** for *x* and *y* respectively.

Statistical Independence

- Random variables x and y are said to be statistically independent if and only if $P(x, y) = P_x(x) P_y(y)$

Covariance

- Cov(x, y), often denoted σ_{xy}
- Covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} = E[(x-\mu)(x-\mu)^t]$

Conditional Density

- $P(x|y) =$

- Similarly, we may write the Bayes Rule in terms of densities.

How about continuous random variables?

- Instead of $P(x)$, we have the probability density function (PDF) $p(x)$

- Some properties of $p(x)$:

- The cumulative distribution function (CDF) $F(x)$:

Continuous Random Variables

- **Mean, variance, etc., are similarly defined, via integrals.**

- **Joint PDF $p(x,y)$ of two variables**
  - Marginal PDFs for x and y
  - If $x \sim p_x(x)$ and $y \sim p_y(y)$ are independent $p(x,y) =$

Continuous Random Variables

- Conditional PDF $p(x|y)$
- Bayes rule for PDF:

Review of Mathematical Foundations

Common Densities

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Discuss common densities useful for machine learning application

Common Distributions

- Uniform Distribution
- Normal (Gaussian) Distribution

The Uniform Distribution, *U*(*a*, *b*)

1-D example, with PDF

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & o.w. \end{cases}$$

The Uniform Distribution, *U*(*a*, *b*)

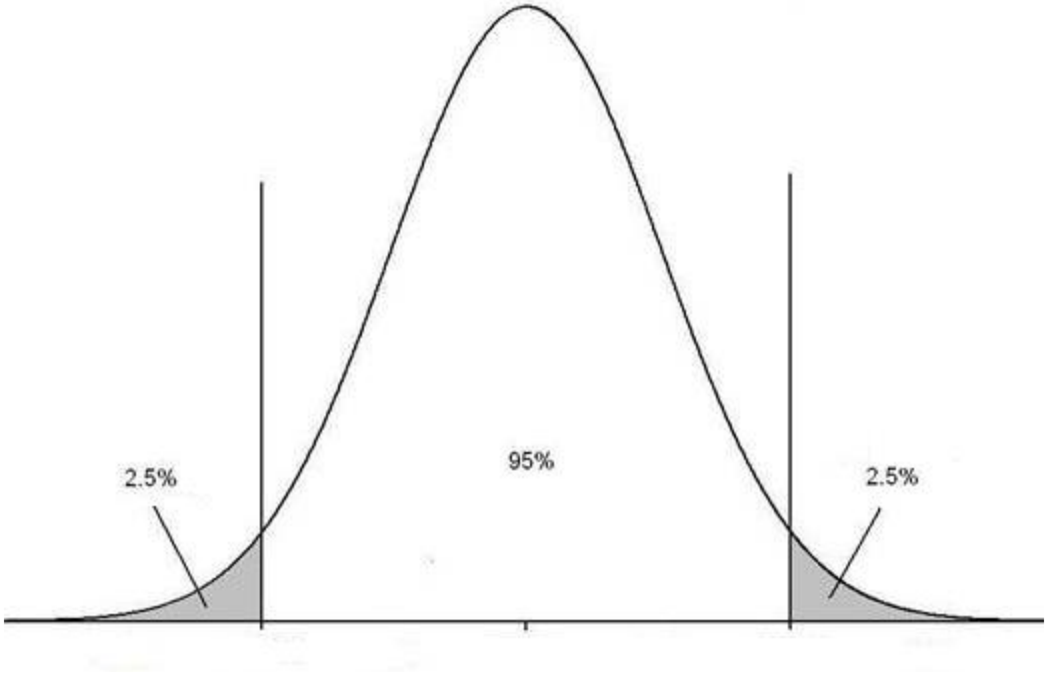
What is the CDF of *p*(*x*)?

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & o.w. \end{cases}$$

The Normal Distribution, *N*(μ, σ^2)

1-D example, with PDF

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

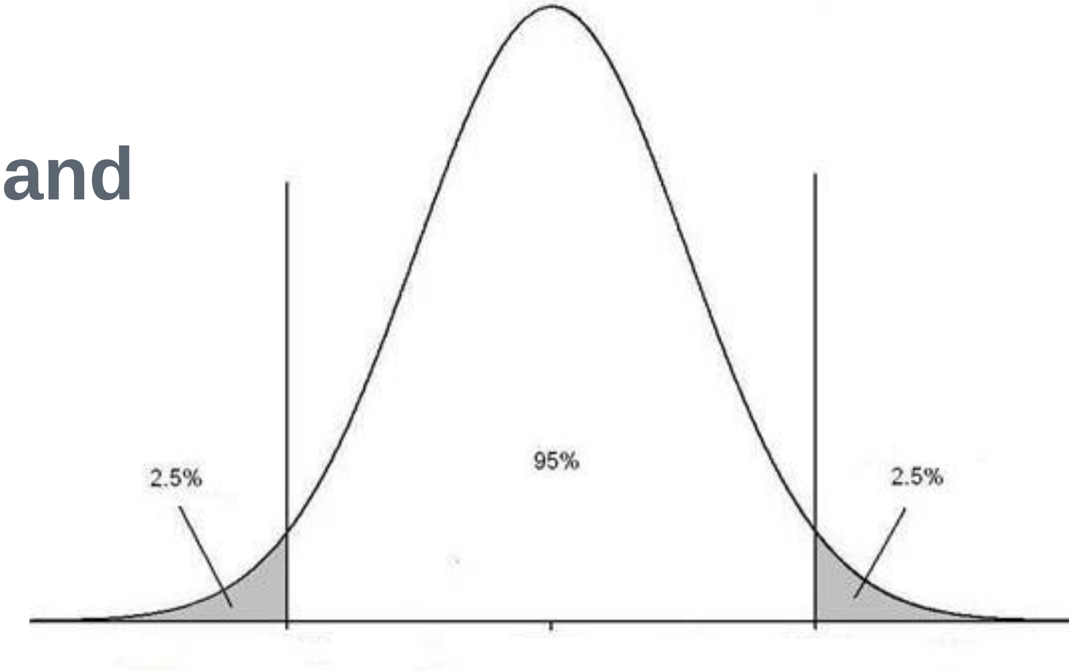


The Normal Distribution, $N(\mu, \sigma^2)$

- 1-D example, with PDF

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- What is the mean and variance ?



Standardized Normal Distribution

- 1-D example, with PDF

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- What is the CDF?
- The error function

$$\operatorname{erf}(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-x^2} dx$$

CDF for General Normal Distribution

What is the CDF for *N*(μ, σ^2)?

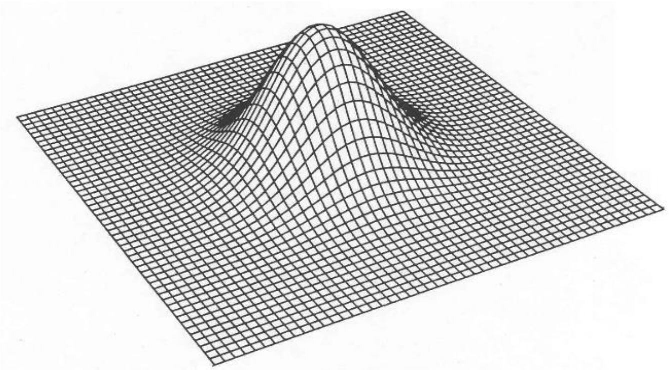
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Multivariate Normal Distribution

- *d*-dimensional vector x is said to be of multivariate normal distribution if its PDF is of the form

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu)]$$

- Visualization of a 2-*d* example



Whitening Transformation

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu)]$$

- **Given some data x distributed according to the above density, we may apply some transformation to x, so that the covariance matrix of the transformed data is diagonal.**
  - The transformation can be formed by the eigenvectors of Σ