

# Probability & Information Theory

Shan-Hung Wu  
*shwu@cs.nthu.edu.tw*

Department of Computer Science,  
National Tsing Hua University, Taiwan

Large-Scale ML, Fall 2016

# Outline

- 1 Random Variables & Probability Distributions
- 2 Multivariate & Derived Random Variables
- 3 Bayes' Rule & Statistics
- 4 Application: Principal Components Analysis
- 5 Technical Details of Random Variables
- 6 Common Probability Distributions
- 7 Common Parametrizing Functions
- 8 Information Theory
- 9 Application: Decision Trees & Random Forest

# Outline

- 1 Random Variables & Probability Distributions
- 2 Multivariate & Derived Random Variables
- 3 Bayes' Rule & Statistics
- 4 Application: Principal Components Analysis
- 5 Technical Details of Random Variables
- 6 Common Probability Distributions
- 7 Common Parametrizing Functions
- 8 Information Theory
- 9 Application: Decision Trees & Random Forest

# Random Variables

- A **random variable**  $x$  is a variable that can take on different values randomly
  - E.g.,  $\Pr(x = x_1) = 0.1$ ,  $\Pr(x = x_2) = 0.3$ , etc.
  - Technically,  $x$  is a function that maps events to a real values
- Must be coupled with a **probability distribution**  $P$  that specifies how likely each value is
  - $x \sim P(\theta)$  means “ $x$  has distribution  $P$  parametrized by  $\theta$ ”

# Probability Mass and Density Functions

- If  $x$  is discrete,  $P(x = x)$  denotes a *probability mass function*  
 $P_x(x) = \Pr(x = x)$ 
  - E.g., the output of a fair dice has discrete uniform distribution with  
 $P(x) = 1/6$

# Probability Mass and Density Functions

- If  $x$  is discrete,  $P(x = x)$  denotes a *probability mass function*  
 $P_x(x) = \Pr(x = x)$ 
  - E.g., the output of a fair dice has discrete uniform distribution with  
 $P(x) = 1/6$
- If  $x$  is continuous,  $P(x = x)$  denotes a *probability density function*  
 $p_x(x)$

# Probability Mass and Density Functions

- If  $x$  is discrete,  $P(x = x)$  denotes a *probability mass function*  
 $P_x(x) = \Pr(x = x)$ 
  - E.g., the output of a fair dice has discrete uniform distribution with  
 $P(x) = 1/6$
- If  $x$  is continuous,  $P(x = x)$  denotes a *probability density function*  
 $p_x(x)$ 
  - Is  $p_x(x)$  a probability?

# Probability Mass and Density Functions

- If  $x$  is discrete,  $P(x = x)$  denotes a *probability mass function*  
 $P_x(x) = \Pr(x = x)$ 
  - E.g., the output of a fair dice has discrete uniform distribution with  
 $P(x) = 1/6$
- If  $x$  is continuous,  $P(x = x)$  denotes a *probability density function*  
 $p_x(x)$ 
  - Is  $p_x(x)$  a probability? **No**, it is “rate of increase in probability at  $x$ ”

$$\Pr(a \leq x \leq b) = \int_{[a,b]} p(x) dx$$



# Probability Mass and Density Functions

- If  $x$  is discrete,  $P(x = x)$  denotes a *probability mass function*  
 $P_x(x) = \Pr(x = x)$ 
  - E.g., the output of a fair dice has discrete uniform distribution with  
 $P(x) = 1/6$
- If  $x$  is continuous,  $P(x = x)$  denotes a *probability density function*  
 $p_x(x)$ 
  - Is  $p_x(x)$  a probability? **No**, it is “rate of increase in probability at  $x$ ”

$$\Pr(a \leq x \leq b) = \int_{[a,b]} p(x) dx$$

- $p_x(x)$  can be greater than 1
- E.g., a continuous uniform distribution within  $[a, b]$  has  $p(x) = 1/b-a$  if  $x \in [a, b]$ ; 0 otherwise

# Marginal Probability

- Consider a probability distribution over a set of variables, e.g.,  $P(x, y)$
- The probability distribution over the subset of random variables called the *marginal probability* distribution:

$$P(x = x) = \sum_y P(x, y) \quad \text{or} \quad \int p(x, y) dy$$

- Also called the sum rule of probability

# Conditional Probability

- Conditional density function:

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

- Defined only when  $P(y = y) > 0$

# Conditional Probability

- Conditional density function:

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

- Defined only when  $P(y = y) > 0$
- Product rule of probability:

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

- E.g.,  $P(a, b, c) = P(a | b, c)P(b | c)P(c)$

# Independence and Conditional Independence

- We say random variables  $x$  is *independent* with  $y$  iff

$$P(x|y) = P(x)$$

- Implies  $P(x, y) = P(x)P(y)$
- Denoted by  $x \perp y$

# Independence and Conditional Independence

- We say random variables  $x$  is *independent* with  $y$  iff

$$P(x|y) = P(x)$$

- Implies  $P(x, y) = P(x)P(y)$
  - Denoted by  $x \perp y$
- We say random variables  $x$  is *conditionally independent* with  $y$  given  $z$  iff

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

- Implies  $P(x, y|z) = P(x|z)P(y|z)$
  - Denoted by  $x \perp y|z$

# Expectation

- The *expectation* (or *expected value* or *mean*) of some function  $f$  with respect to  $x$  is the “average” value that  $f$  takes on:<sup>1</sup>

$$E_{x \sim P}[f(x)] = \sum_x P_x(x)f(x) \text{ or } \int p_x(x)f(x)dx = \mu_{f(x)}$$

---

<sup>1</sup>The bracket  $[\cdot]$  here is used to distinguish the parentheses inside and has nothing to do with functionals.

# Expectation

- The *expectation* (or *expected value* or *mean*) of some function  $f$  with respect to  $x$  is the “average” value that  $f$  takes on:<sup>1</sup>

$$E_{x \sim P}[f(x)] = \sum_x P_x(x)f(x) \text{ or } \int p_x(x)f(x)dx = \mu_{f(x)}$$

- Expectation is linear:  $E[af(x) + b] = aE[f(x)] + b$  for deterministic  $a$  and  $b$

---

<sup>1</sup>The bracket  $[\cdot]$  here is used to distinguish the parentheses inside and has nothing to do with functionals.



# Expectation

- The *expectation* (or *expected value* or *mean*) of some function  $f$  with respect to  $x$  is the “average” value that  $f$  takes on:<sup>1</sup>

$$E_{x \sim P}[f(x)] = \sum_x P_x(x)f(x) \text{ or } \int p_x(x)f(x)dx = \mu_{f(x)}$$

- Expectation is linear:  $E[af(x) + b] = aE[f(x)] + b$  for deterministic  $a$  and  $b$
- $E[E[f(x)]] = E[f(x)]$ , as  $E[f(x)]$  is deterministic

---

<sup>1</sup>The bracket  $[\cdot]$  here is used to distinguish the parentheses inside and has nothing to do with functionals.

# Expectation over Multiple Variables

- Defined over the join probability distribution, e.g.,

$$E[f(x,y)] = \sum_{x,y} P_{x,y}(x,y)f(x,y) \text{ or } \int_{x,y} p_{x,y}(x,y)f(x,y)dxdy$$

# Expectation over Multiple Variables

- Defined over the join probability distribution, e.g.,

$$E[f(x,y)] = \sum_{x,y} P_{x,y}(x,y)f(x,y) \text{ or } \int_{x,y} p_{x,y}(x,y)f(x,y)dxdy$$

- $E[f(x) | y = y] = \int p_{x|y}(x|y)f(x)dx$  is called the *conditional expectation*
- $E[f(x)g(y)] = E[f(x)]E[g(y)]$  if  $x$  and  $y$  are independent [Proof]

# Variance

- The *variance* measures how much the values of  $f$  deviate from its expected value when seeing different values of  $x$ :

$$\text{Var}[f(x)] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \sigma_{f(x)}^2$$

- $\sigma_{f(x)}$  is called the *standard deviation*

# Variance

- The *variance* measures how much the values of  $f$  deviate from its expected value when seeing different values of  $x$ :

$$\text{Var}[f(x)] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \sigma_{f(x)}^2$$

- $\sigma_{f(x)}$  is called the *standard deviation*
- $\text{Var}[af(x) + b] = a^2 \text{Var}[f(x)]$  for deterministic  $a$  and  $b$  [Proof]

# Covariance I

- *Covariance* gives some sense of how much two values are *linearly* related to each other

$$\text{Cov}[f(x), g(y)] = E[(f(x) - E[f(x)])(g(y) - E[g(y)])]$$

- If sign positive, both variables tend to take on high values simultaneously
- If sign negative, one variable tend to take on high value while the other taking on low one

# Covariance I

- **Covariance** gives some sense of how much two values are **linearly** related to each other

$$\text{Cov}[f(x), g(y)] = E[(f(x) - E[f(x)])(g(y) - E[g(y)])]$$

- If sign positive, both variables tend to take on high values simultaneously
- If sign negative, one variable tend to take on high value while the other taking on low one
- If  $x$  and  $y$  are independent, then  $\text{Cov}(x, y) = 0$  [Proof]
  - The converse is **not** true as  $X$  and  $Y$  may be related in a nonlinear way
  - E.g.,  $y = \sin(x)$  and  $x \sim \text{Uniform}(-\pi, \pi)$

# Covariance II

- $\text{Var}(ax + by) = a^2\text{Var}(x) + b^2\text{Var}(y) + 2ab\text{Cov}(x, y)$  [Proof]



# Covariance II

- $\text{Var}(ax + by) = a^2\text{Var}(x) + b^2\text{Var}(y) + 2ab\text{Cov}(x, y)$  [Proof]
  - $\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y)$  if  $x$  and  $y$  are independent
- $\text{Cov}(ax + b, cy + d) = ac\text{Cov}(x, y)$  [Proof]

# Covariance II

- $\text{Var}(ax + by) = a^2\text{Var}(x) + b^2\text{Var}(y) + 2ab\text{Cov}(x, y)$  [Proof]
  - $\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y)$  if  $x$  and  $y$  are independent
- $\text{Cov}(ax + b, cy + d) = ac\text{Cov}(x, y)$  [Proof]
- $\text{Cov}(ax + by, cw + dv) =$   
 $ac\text{Cov}(x, w) + ad\text{Cov}(x, v) + bc\text{Cov}(y, w) + bd\text{Cov}(y, v)$  [Proof]

# Outline

- 1 Random Variables & Probability Distributions
- 2 Multivariate & Derived Random Variables**
- 3 Bayes' Rule & Statistics
- 4 Application: Principal Components Analysis
- 5 Technical Details of Random Variables
- 6 Common Probability Distributions
- 7 Common Parametrizing Functions
- 8 Information Theory
- 9 Application: Decision Trees & Random Forest

# Multivariate Random Variables I

- A multivariate random variable is denoted by  $\mathbf{x} = [x_1, \dots, x_d]^\top$ 
  - Normally,  $x_i$ 's (*attributes* or *variables* or *features*) are dependent with each other
  - $P(\mathbf{x})$  is a joint distribution of  $x_1, \dots, x_d$

# Multivariate Random Variables I

- A multivariate random variable is denoted by  $\mathbf{x} = [x_1, \dots, x_d]^\top$ 
  - Normally,  $x_i$ 's (**attributes** or **variables** or **features**) are dependent with each other
  - $P(\mathbf{x})$  is a joint distribution of  $x_1, \dots, x_d$
- The **mean** of  $\mathbf{x}$  is defined as  $\mu_{\mathbf{x}} = E(\mathbf{x}) = [\mu_{x_1}, \dots, \mu_{x_d}]^\top$

# Multivariate Random Variables I

- A multivariate random variable is denoted by  $\mathbf{x} = [x_1, \dots, x_d]^\top$ 
  - Normally,  $x_i$ 's (**attributes** or **variables** or **features**) are dependent with each other
  - $P(\mathbf{x})$  is a joint distribution of  $x_1, \dots, x_d$
- The **mean** of  $\mathbf{x}$  is defined as  $\mu_{\mathbf{x}} = E(\mathbf{x}) = [\mu_{x_1}, \dots, \mu_{x_d}]^\top$
- The **covariance matrix** of  $\mathbf{x}$  is defined as:

$$\Sigma_{\mathbf{x}} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1, x_2} & \cdots & \sigma_{x_1, x_d} \\ \sigma_{x_2, x_1} & \sigma_{x_2}^2 & \cdots & \sigma_{x_2, x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_d, x_1} & \sigma_{x_d, x_2} & \cdots & \sigma_{x_d}^2 \end{bmatrix}$$

- $\sigma_{x_i, x_j} = \text{Cov}(x_i, x_j) = E[(x_i - \mu_{x_i})(x_j - \mu_{x_j})] = E(x_i x_j) - \mu_{x_i} \mu_{x_j}$

# Multivariate Random Variables I

- A multivariate random variable is denoted by  $\mathbf{x} = [x_1, \dots, x_d]^\top$ 
  - Normally,  $x_i$ 's (**attributes** or **variables** or **features**) are dependent with each other
  - $P(\mathbf{x})$  is a joint distribution of  $x_1, \dots, x_d$
- The **mean** of  $\mathbf{x}$  is defined as  $\mu_{\mathbf{x}} = E(\mathbf{x}) = [\mu_{x_1}, \dots, \mu_{x_d}]^\top$
- The **covariance matrix** of  $\mathbf{x}$  is defined as:

$$\Sigma_{\mathbf{x}} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1, x_2} & \cdots & \sigma_{x_1, x_d} \\ \sigma_{x_2, x_1} & \sigma_{x_2}^2 & \cdots & \sigma_{x_2, x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_d, x_1} & \sigma_{x_d, x_2} & \cdots & \sigma_{x_d}^2 \end{bmatrix}$$

- $\sigma_{x_i, x_j} = \text{Cov}(x_i, x_j) = E[(x_i - \mu_{x_i})(x_j - \mu_{x_j})] = E(x_i x_j) - \mu_{x_i} \mu_{x_j}$
- $\Sigma_{\mathbf{x}} = \text{Cov}(\mathbf{x}) = E[(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})^\top] = E(\mathbf{x} \mathbf{x}^\top) - \mu_{\mathbf{x}} \mu_{\mathbf{x}}^\top$

# Multivariate Random Variables II

- $\Sigma_{\mathbf{x}}$  is always symmetric



# Multivariate Random Variables II

- $\Sigma_{\mathbf{x}}$  is always symmetric
- $\Sigma_{\mathbf{x}}$  is always positive semidefinite [Homework]

# Multivariate Random Variables II

- $\Sigma_{\mathbf{x}}$  is always symmetric
- $\Sigma_{\mathbf{x}}$  is always positive semidefinite [Homework]
- $\Sigma_{\mathbf{x}}$  is nonsingular iff it is positive definite

# Multivariate Random Variables II

- $\Sigma_{\mathbf{x}}$  is always symmetric
- $\Sigma_{\mathbf{x}}$  is always positive semidefinite [Homework]
- $\Sigma_{\mathbf{x}}$  is nonsingular iff it is positive definite
- $\Sigma_{\mathbf{x}}$  is singular implies that  $\mathbf{x}$  has either:
  - Deterministic/independent/non-linearly dependent attributes causing zero rows, or
  - Redundant attributes causing linear dependency between rows

# Derived Random Variables

- Let  $y = f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^\top \mathbf{x}$  a random variable transformed from  $\mathbf{x}$
- $\mu_y = E(\mathbf{w}^\top \mathbf{x}) = \mathbf{w}^\top E(\mathbf{x}) = \mathbf{w}^\top \mu_{\mathbf{x}}$
- $\sigma_y^2 = \mathbf{w}^\top \Sigma_{\mathbf{x}} \mathbf{w}$  [Homework]

# Outline

- 1 Random Variables & Probability Distributions
- 2 Multivariate & Derived Random Variables
- 3 Bayes' Rule & Statistics**
- 4 Application: Principal Components Analysis
- 5 Technical Details of Random Variables
- 6 Common Probability Distributions
- 7 Common Parametrizing Functions
- 8 Information Theory
- 9 Application: Decision Trees & Random Forest

# What Does $\Pr(\mathbf{x} = x)$ Mean?

# What Does $\Pr(x = x)$ Mean?

- ① *Bayesian probability*: it's a degree of belief or qualitative levels of certainty

# What Does $\Pr(x = x)$ Mean?

- ① **Bayesian probability**: it's a degree of belief or qualitative levels of certainty
- ② **Frequentist probability**: if we can draw samples of  $x$ , then the proportion of frequency of samples having the value  $x$  is equal to  $\Pr(x = x)$



# Bayes' Rule

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

- Bayes' Rule is so important in statistics (and ML as well) such that each term has a name:

$$\textit{posterior of } y = \frac{(\textit{likelihood of } y) \times (\textit{prior of } y)}{\textit{evidence}}$$

X症狀    Y得到的病

P(Y|X) 流鼻涕的話是感冒的機率

P(X) 流鼻涕的機率

P(Y) 感冒的機率

P(X|Y) 感冒的的話流鼻涕的機率

# Bayes' Rule

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

- Bayes' Rule is so important in statistics (and ML as well) such that each term has a name:

$$\textit{posterior of } y = \frac{(\textit{likelihood of } y) \times (\textit{prior of } y)}{\textit{evidence}}$$

- Why is it so important?

# Bayes' Rule

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

- Bayes' Rule is so important in statistics (and ML as well) such that each term has a name:

$$\textit{posterior of } y = \frac{(\textit{likelihood of } y) \times (\textit{prior of } y)}{\textit{evidence}}$$

- Why is it so important?
- E.g., a doctor diagnoses you as having a disease by letting  $x$  be “symptom” and  $y$  be “disease”
  - $P(x|y)$  and  $P(y)$  may be estimated from sample frequencies more easily

# Point Estimation

- *Point estimation* is the attempt to estimate some fixed but unknown quantity  $\theta$  of a random variable by using sample data

# Point Estimation

- **Point estimation** is the attempt to estimate some fixed but unknown quantity  $\theta$  of a random variable by using sample data
- Let  $\{x^{(1)}, \dots, x^{(n)}\}$  be a set of  $n$  independent and identically distributed (**i.i.d.**) samples of a random variable  $x$ , a **point estimator** or **statistic** is a function of the data:

$$\hat{\theta}_n = g(x^{(1)}, \dots, x^{(n)})$$

- $\hat{\theta}_n$  is called the **estimate** of  $\theta$

# Sample Mean and Covariance

- Given  $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}]^\top \in \mathbb{R}^{n \times d}$  the i.i.d samples, what are the estimates of the mean and covariance of  $\mathbf{x}$ ?

# Sample Mean and Covariance

- Given  $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}]^\top \in \mathbb{R}^{n \times d}$  the i.i.d samples, what are the estimates of the mean and covariance of  $\mathbf{x}$ ?
- A sample mean:

$$\hat{\mu}_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^{(i)}$$

# Sample Mean and Covariance

- Given  $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}]^\top \in \mathbb{R}^{n \times d}$  the i.i.d samples, what are the estimates of the mean and covariance of  $\mathbf{x}$ ?
- A sample mean:

$$\hat{\mu}_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^{(i)}$$

- A sample covariance matrix:

$$\hat{\Sigma}_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}^{(i)} - \hat{\mu}_{\mathbf{x}})(\mathbf{x}^{(i)} - \hat{\mu}_{\mathbf{x}})^\top$$

- $\hat{\sigma}_{x_i, x_j}^2 = \frac{1}{n} \sum_{s=1}^n (x_i^{(s)} - \hat{\mu}_{x_i})(x_j^{(s)} - \hat{\mu}_{x_j})$



# Sample Mean and Covariance

- Given  $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}]^\top \in \mathbb{R}^{n \times d}$  the i.i.d samples, what are the estimates of the mean and covariance of  $\mathbf{x}$ ?
- A sample mean:

$$\hat{\mu}_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^{(i)}$$

- A sample covariance matrix:

$$\hat{\Sigma}_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}^{(i)} - \hat{\mu}_{\mathbf{x}})(\mathbf{x}^{(i)} - \hat{\mu}_{\mathbf{x}})^\top$$

- $\hat{\sigma}_{x_i, x_j}^2 = \frac{1}{n} \sum_{s=1}^n (x_i^{(s)} - \hat{\mu}_{x_i})(x_j^{(s)} - \hat{\mu}_{x_j})$
- If each  $\mathbf{x}^{(i)}$  is centered (by subtracting  $\hat{\mu}_{\mathbf{x}}$  first), then  $\hat{\Sigma}_{\mathbf{x}} = \frac{1}{n} \mathbf{X}^\top \mathbf{X}$

# Outline

- 1 Random Variables & Probability Distributions
- 2 Multivariate & Derived Random Variables
- 3 Bayes' Rule & Statistics
- 4 Application: Principal Components Analysis**
- 5 Technical Details of Random Variables
- 6 Common Probability Distributions
- 7 Common Parametrizing Functions
- 8 Information Theory
- 9 Application: Decision Trees & Random Forest

# Principal Components Analysis (PCA) I

<https://medium.com/@chih.sheng.huang821/%E6%A9%9F%E5%99%A8-%E7%B5%B1%E8%A8%88%E5%AD%B8%E7%BF%92-%E4%B8%BB%E6%88%90%E5%88%86%E5%88%86%E6%9E%90-principle-component-analysis-pca-58229cd26e71>

- Give a collection of data points  $\mathbb{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ , where  $\mathbf{x}^{(i)} \in \mathbb{R}^D$
- Suppose we want to lossily compress  $\mathbb{X}$ , i.e., to find a function  $f$  such that  $f(\mathbf{x}^{(i)}) = \mathbf{z}^{(i)} \in \mathbb{R}^K$ , where  $K < D$
- How to keep the maximum info in  $\mathbb{X}$ ?

# Principal Components Analysis (PCA) II

- Let  $\mathbf{x}^{(i)}$ 's be i.i.d. samples of a random variable  $\mathbf{x}$
- Let  $f$  be linear, i.e.,  $f(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}$  for some  $\mathbf{W} \in \mathbb{R}^{D \times K}$

# Principal Components Analysis (PCA) II

- Let  $\mathbf{x}^{(i)}$ 's be i.i.d. samples of a random variable  $\mathbf{x}$
- Let  $f$  be linear, i.e.,  $f(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}$  for some  $\mathbf{W} \in \mathbb{R}^{D \times K}$
- **Principal Component Analysis (PCA)** finds  $K$  orthonormal vectors  $\mathbf{W} = [\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}]$  such that the transformed variable  $\mathbf{z} = \mathbf{W}^\top \mathbf{x}$  has the most “spread out” attributes, i.e., each attribute  $z_j = \mathbf{w}^{(j)\top} \mathbf{x}$  has the maximum variance  $\text{Var}(z_j)$ 
  - $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}$  are called the **principle components**

# Principal Components Analysis (PCA) II

- Let  $\mathbf{x}^{(i)}$ 's be i.i.d. samples of a random variable  $\mathbf{x}$
- Let  $f$  be linear, i.e.,  $f(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}$  for some  $\mathbf{W} \in \mathbb{R}^{D \times K}$
- **Principal Component Analysis (PCA)** finds  $K$  orthonormal vectors  $\mathbf{W} = [\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}]$  such that the transformed variable  $\mathbf{z} = \mathbf{W}^\top \mathbf{x}$  has the most “spread out” attributes, i.e., each attribute  $z_j = \mathbf{w}^{(j)\top} \mathbf{x}$  has the maximum variance  $\text{Var}(z_j)$ 
  - $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}$  are called the **principle components**
- Why  $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}$  need to be orthogonal with each other?

# Principal Components Analysis (PCA) II

- Let  $\mathbf{x}^{(i)}$ 's be i.i.d. samples of a random variable  $\mathbf{x}$
- Let  $f$  be linear, i.e.,  $f(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}$  for some  $\mathbf{W} \in \mathbb{R}^{D \times K}$
- **Principal Component Analysis (PCA)** finds  $K$  orthonormal vectors  $\mathbf{W} = [\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}]$  such that the transformed variable  $\mathbf{z} = \mathbf{W}^\top \mathbf{x}$  has the most “spread out” attributes, i.e., each attribute  $z_j = \mathbf{w}^{(j)\top} \mathbf{x}$  has the maximum variance  $\text{Var}(z_j)$ 
  - $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}$  are called the **principle components**
- Why  $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}$  need to be orthogonal with each other?
  - Each  $\mathbf{w}^{(j)}$  keeps information that cannot be explained by others, so together they preserve the most info

# Principal Components Analysis (PCA) II

- Let  $\mathbf{x}^{(i)}$ 's be i.i.d. samples of a random variable  $\mathbf{x}$
- Let  $f$  be linear, i.e.,  $f(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}$  for some  $\mathbf{W} \in \mathbb{R}^{D \times K}$
- Principal Component Analysis (PCA)** finds  $K$  orthonormal vectors  $\mathbf{W} = [\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}]$  such that the transformed variable  $\mathbf{z} = \mathbf{W}^\top \mathbf{x}$  has the most “spread out” attributes, i.e., each attribute  $z_j = \mathbf{w}^{(j)\top} \mathbf{x}$  has the maximum variance  $\text{Var}(z_j)$ 
  - $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}$  are called the **principle components**
- Why  $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}$  need to be orthogonal with each other?
  - Each  $\mathbf{w}^{(j)}$  keeps information that cannot be explained by others, so together they preserve the most info
- Why  $\|\mathbf{w}^{(j)}\| = 1$  for all  $j$ ?

orthogonal

代表變換後會有多餘

的維度

那W2就是多

的維度



# Principal Components Analysis (PCA) II

- Let  $\mathbf{x}^{(i)}$ 's be i.i.d. samples of a random variable  $\mathbf{x}$
- Let  $f$  be linear, i.e.,  $f(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}$  for some  $\mathbf{W} \in \mathbb{R}^{D \times K}$
- Principal Component Analysis (PCA)** finds  $K$  orthonormal vectors  $\mathbf{W} = [\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}]$  such that the transformed variable  $\mathbf{z} = \mathbf{W}^\top \mathbf{x}$  has the most “spread out” attributes, i.e., each attribute  $z_j = \mathbf{w}^{(j)\top} \mathbf{x}$  has the maximum variance  $\text{Var}(z_j)$ 
  - $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}$  are called the **principle components**
- Why  $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}$  need to be orthogonal with each other?
  - Each  $\mathbf{w}^{(j)}$  keeps information that cannot be explained by others, so together they preserve the most info
- Why  $\|\mathbf{w}^{(j)}\| = 1$  for all  $j$ ?
  - Only directions matter—we don't want to maximize  $\text{Var}(z_j)$  by finding a long  $\mathbf{w}^{(j)}$   
為了希望 $\mathbf{W}$ 的長度不會影響到 $\text{Var}$ 的判斷  
否則我把 $\mathbf{W}$ 用到超大  $\text{Var}$ 就會因此變大

# Solving $W$ I

- For simplicity, let's consider  $K = 1$  first
- How to evaluate  $\text{Var}(z_1)$ ?

# Solving $W$ I

- For simplicity, let's consider  $K = 1$  first
- How to evaluate  $\text{Var}(z_1)$ ?
  - Recall that  $z_1 = \mathbf{w}^{(1)\top} \mathbf{x}$  implies  $\sigma_{z_1}^2 = \mathbf{w}^{(1)\top} \Sigma_{\mathbf{x}} \mathbf{w}^{(1)}$  [Homework]
  - How to get  $\Sigma_{\mathbf{x}}$ ?

# Solving $W$ I

- For simplicity, let's consider  $K = 1$  first
- How to evaluate  $\text{Var}(z_1)$ ?
  - Recall that  $z_1 = \mathbf{w}^{(1)\top} \mathbf{x}$  implies  $\sigma_{z_1}^2 = \mathbf{w}^{(1)\top} \Sigma_{\mathbf{x}} \mathbf{w}^{(1)}$  [Homework]
  - How to get  $\Sigma_{\mathbf{x}}$ ?
  - An estimate:  $\hat{\Sigma}_{\mathbf{x}} = \frac{1}{N} \mathbf{X}^\top \mathbf{X}$  (assuming  $\mathbf{x}^{(i)}$ 's are centered first)

# Solving $W$ I

- For simplicity, let's consider  $K = 1$  first
- How to evaluate  $\text{Var}(z_1)$ ?
  - Recall that  $z_1 = \mathbf{w}^{(1)\top} \mathbf{x}$  implies  $\sigma_{z_1}^2 = \mathbf{w}^{(1)\top} \Sigma_{\mathbf{x}} \mathbf{w}^{(1)}$  [Homework]
  - How to get  $\Sigma_{\mathbf{x}}$ ?
  - An estimate:  $\hat{\Sigma}_{\mathbf{x}} = \frac{1}{N} \mathbf{X}^\top \mathbf{X}$  (assuming  $\mathbf{x}^{(i)}$ 's are centered first)
- Optimization problem to solve:

$$\arg \max_{\mathbf{w}^{(1)} \in \mathbb{R}^D} \mathbf{w}^{(1)\top} \mathbf{X}^\top \mathbf{X} \mathbf{w}^{(1)}, \text{ subject to } \|\mathbf{w}^{(1)}\| = 1$$

# Solving $W$ I

- For simplicity, let's consider  $K = 1$  first
- How to evaluate  $\text{Var}(z_1)$ ?
  - Recall that  $z_1 = \mathbf{w}^{(1)\top} \mathbf{x}$  implies  $\sigma_{z_1}^2 = \mathbf{w}^{(1)\top} \Sigma_{\mathbf{x}} \mathbf{w}^{(1)}$  [Homework]
  - How to get  $\Sigma_{\mathbf{x}}$ ?
  - An estimate:  $\hat{\Sigma}_{\mathbf{x}} = \frac{1}{N} \mathbf{X}^\top \mathbf{X}$  (assuming  $\mathbf{x}^{(i)}$ 's are centered first)
- Optimization problem to solve:

$$\arg \max_{\mathbf{w}^{(1)} \in \mathbb{R}^D} \mathbf{w}^{(1)\top} \mathbf{X}^\top \mathbf{X} \mathbf{w}^{(1)}, \text{ subject to } \|\mathbf{w}^{(1)}\| = 1$$

- $\mathbf{X}^\top \mathbf{X}$  is symmetric thus can be eigendecomposed

# Solving $W$ I

- For simplicity, let's consider  $K = 1$  first
- How to evaluate  $\text{Var}(z_1)$ ?
  - Recall that  $z_1 = \mathbf{w}^{(1)\top} \mathbf{x}$  implies  $\sigma_{z_1}^2 = \mathbf{w}^{(1)\top} \Sigma_{\mathbf{x}} \mathbf{w}^{(1)}$  [Homework]
  - How to get  $\Sigma_{\mathbf{x}}$ ?
  - An estimate:  $\hat{\Sigma}_{\mathbf{x}} = \frac{1}{N} \mathbf{X}^\top \mathbf{X}$  (assuming  $\mathbf{x}^{(i)}$ 's are centered first)
- Optimization problem to solve:

$$\arg \max_{\mathbf{w}^{(1)} \in \mathbb{R}^D} \mathbf{w}^{(1)\top} \mathbf{X}^\top \mathbf{X} \mathbf{w}^{(1)}, \text{ subject to } \|\mathbf{w}^{(1)}\| = 1$$

- $\mathbf{X}^\top \mathbf{X}$  is symmetric thus can be eigendecomposed
- By Rayleigh's Quotient, the optimal  $\mathbf{w}^{(1)}$  is given by the eigenvector of  $\mathbf{X}^\top \mathbf{X}$  corresponding to the largest eigenvalue

# Solving $W$ II

- Optimization problem for  $\mathbf{w}^{(2)}$ :

$$\arg \max_{\mathbf{w}^{(2)} \in \mathbb{R}^D} \mathbf{w}^{(2)\top} \mathbf{X}^\top \mathbf{X} \mathbf{w}^{(2)}, \text{ subject to } \|\mathbf{w}^{(2)}\| = 1 \text{ and } \mathbf{w}^{(2)\top} \mathbf{w}^{(1)} = 0$$



# Solving $W$ II

- Optimization problem for  $\mathbf{w}^{(2)}$ :

$$\arg \max_{\mathbf{w}^{(2)} \in \mathbb{R}^D} \mathbf{w}^{(2)\top} \mathbf{X}^\top \mathbf{X} \mathbf{w}^{(2)}, \text{ subject to } \|\mathbf{w}^{(2)}\| = 1 \text{ and } \mathbf{w}^{(2)\top} \mathbf{w}^{(1)} = 0$$

- By Rayleigh's Quotient again,  $\mathbf{w}^{(2)}$  is the eigenvector corresponding to the 2-nd largest eigenvalue

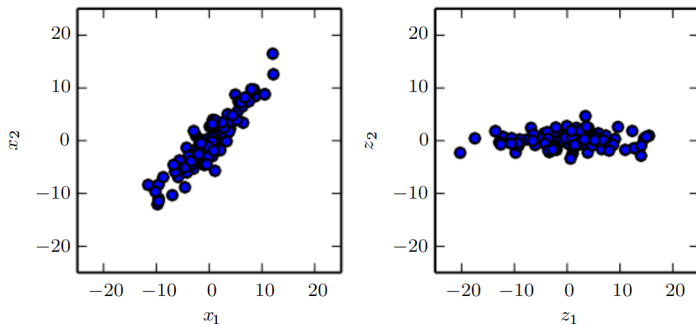
# Solving $W$ II

- Optimization problem for  $\mathbf{w}^{(2)}$ :

$$\arg \max_{\mathbf{w}^{(2)} \in \mathbb{R}^D} \mathbf{w}^{(2)\top} \mathbf{X}^\top \mathbf{X} \mathbf{w}^{(2)}, \text{ subject to } \|\mathbf{w}^{(2)}\| = 1 \text{ and } \mathbf{w}^{(2)\top} \mathbf{w}^{(1)} = 0$$

- By Rayleigh's Quotient again,  $\mathbf{w}^{(2)}$  is the eigenvector corresponding to the 2-nd largest eigenvalue
- For general case where  $K > 1$ , the  $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)}$  are eigenvectors of  $\mathbf{X}^\top \mathbf{X}$  corresponding to the largest  $K$  eigenvalues
  - Proof by induction [Proof]

# Visualization



**Figure:** PCA learns a linear projection that aligns the direction of greatest variance with the axes of the new space. With these new axes, the estimated covariance matrix  $\hat{\Sigma}_{\mathbf{z}} = \mathbf{W}^\top \hat{\Sigma}_{\mathbf{x}} \mathbf{W} \in \mathbb{R}^{K \times K}$  is always diagonal.

# Outline

- 1 Random Variables & Probability Distributions
- 2 Multivariate & Derived Random Variables
- 3 Bayes' Rule & Statistics
- 4 Application: Principal Components Analysis
- 5 Technical Details of Random Variables**
- 6 Common Probability Distributions
- 7 Common Parametrizing Functions
- 8 Information Theory
- 9 Application: Decision Trees & Random Forest

# Sure and Almost Sure Events

- Given a continuous random variable  $x$ , we have  $\Pr(x = x) = 0$  for any value  $x$
- Will the event  $x = x$  occur?

# Sure and Almost Sure Events

- Given a continuous random variable  $x$ , we have  $\Pr(x = x) = 0$  for any value  $x$
- Will the event  $x = x$  occur? *Yes!*
- An event  $\mathbb{A}$  happens *surely* if always occurs
- An event  $\mathbb{A}$  happens *almost surely* if  $\Pr(\mathbb{A}) = 1$  (e.g.,  $\Pr(x \neq x) = 1$ )

# Equality of Random Variables I

## Definition (Equality in Distribution)

Two random variables  $x$  and  $y$  are *equal in distribution* iff  $\Pr(x \leq a) = \Pr(y \leq a)$  for all  $a$ .

## Definition (Almost Sure Equality)

Two random variables  $x$  and  $y$  are *equal almost surely* iff  $\Pr(x = y) = 1$ .

## Definition (Equality)

Two random variables  $x$  and  $y$  are *equal* iff they map the same events to same values.

# Equality of Random Variables II

- What's the difference between the “equality in distribution” and “almost sure equality?”



# Equality of Random Variables II

- What's the difference between the “equality in distribution” and “almost sure equality?”
- Almost sure equality implies equality in distribution, but converse not true

# Equality of Random Variables II

- What's the difference between the “equality in distribution” and “almost sure equality?”
- Almost sure equality implies equality in distribution, but converse not true
- E.g., let  $x$  and  $y$  be binary random variables and  $P_x(0) = P_x(1) = P_y(0) = P_y(1) = 0.5$ 
  - They are equal in distribution
  - But  $\Pr(x = y) = 0.5 \neq 1$

# Convergence of Random Variables I

## Definition (Convergence in Distribution)

A sequence of random variables  $\{x^{(1)}, x^{(2)}, \dots\}$  *converges in distribution* to  $x$  iff  $\lim_{n \rightarrow \infty} P(x^{(n)} = x) = P(x = x)$

## Definition (Convergence in Probability)

A sequence of random variables  $\{x^{(1)}, x^{(2)}, \dots\}$  *converges in probability* to  $x$  iff for any  $\varepsilon > 0$ ,  $\lim_{n \rightarrow \infty} \Pr(|x^{(n)} - x| < \varepsilon) = 1$ .

## Definition (Almost Sure Convergence)

A sequence of random variables  $\{x^{(1)}, x^{(2)}, \dots\}$  *converges almost surely* to  $x$  iff  $\Pr(\lim_{n \rightarrow \infty} x^{(n)} = x) = 1$ .

# Convergence of Random Variables II

- What's the difference between the convergence “in probability” and “almost surely?”

# Convergence of Random Variables II

- What's the difference between the convergence “in probability” and “almost surely?”
- Almost sure convergence implies convergence in probability, but converse not true

# Convergence of Random Variables II

- What's the difference between the convergence “in probability” and “almost surely?”
- Almost sure convergence implies convergence in probability, but converse not true
- $\lim_{n \rightarrow \infty} \Pr(|x^{(n)} - x| < \varepsilon) = 1$  leaves open the possibility that  $|x^{(n)} - x| > \varepsilon$  happens an infinite number of times
- $\Pr(\lim_{n \rightarrow \infty} x^{(n)} = x) = 1$  guarantees that  $|x^{(n)} - x| > \varepsilon$  almost surely will not occur

# Distribution of Derived Variables I

- Suppose  $y = f(x)$  and  $f^{-1}$  exists, does  $P(y = y) = P(x = f^{-1}(y))$  always hold?

# Distribution of Derived Variables I

- Suppose  $y = f(x)$  and  $f^{-1}$  exists, does  $P(y = y) = P(x = f^{-1}(y))$  always hold? **No**, when  $x$  and  $y$  are continuous
- Suppose  $x \sim \text{Uniform}(0, 1)$  is continuous and  $p(x) = c$  for  $x \in (0, 1)$
- Let  $y = x/2 \sim \text{Uniform}(0, 1/2)$
- If  $p_y(y) = p_x(2y)$ , then

$$\int_{y=0}^{1/2} p_y(y) dy = \int_{y=0}^{1/2} c \cdot dy = \frac{1}{2} \neq 1$$

- Violates the axiom of probability



# Distribution of Derived Variables II

- Recall that  $\Pr(y = y) = p_y(y)dy$  and  $\Pr(x = x) = p_x(x)dx$

## Distribution of Derived Variables II

- Recall that  $\Pr(y = y) = p_y(y)dy$  and  $\Pr(x = x) = p_x(x)dx$
- Since  $f$  may distort space, we need to ensure that

$$|p_y(f(x))dy| = |p_x(x)dx|$$

- We have

$$p_y(y) = p_x(f^{-1}(y)) \left| \frac{\partial f^{-1}(y)}{\partial y} \right| \quad (\text{or } p_x(x) = p_y(f(x)) \left| \frac{\partial f(x)}{\partial x} \right|)$$

- In previous example:  $p_y(y) = \mathbf{2} \cdot p_x(2y)$

## Distribution of Derived Variables II

- Recall that  $\Pr(y = y) = p_y(y)dy$  and  $\Pr(x = x) = p_x(x)dx$
- Since  $f$  may distort space, we need to ensure that

$$|p_y(f(x))dy| = |p_x(x)dx|$$

- We have

$$p_y(y) = p_x(f^{-1}(y)) \left| \frac{\partial f^{-1}(y)}{\partial y} \right| \quad (\text{or } p_x(x) = p_y(f(x)) \left| \frac{\partial f(x)}{\partial x} \right|)$$

- In previous example:  $p_y(y) = 2 \cdot p_x(2y)$
- In multivariate case, we have

$$p_y(\mathbf{y}) = p_x(\mathbf{f}^{-1}(\mathbf{y})) |\det(\mathbf{J}(\mathbf{f}^{-1})(\mathbf{y}))|,$$

where  $\mathbf{J}(\mathbf{f}^{-1})(\mathbf{y})$  is the Jacobian matrix of  $\mathbf{f}^{-1}$  at input  $\mathbf{y}$

- $\mathbf{J}(\mathbf{f}^{-1})(\mathbf{y})_{i,j} = \partial f_i^{-1}(\mathbf{y}) / \partial y_j$

# Outline

- 1 Random Variables & Probability Distributions
- 2 Multivariate & Derived Random Variables
- 3 Bayes' Rule & Statistics
- 4 Application: Principal Components Analysis
- 5 Technical Details of Random Variables
- 6 Common Probability Distributions**
- 7 Common Parametrizing Functions
- 8 Information Theory
- 9 Application: Decision Trees & Random Forest

# Random Experiments

- The value of a random variable  $x$  can be think of as the outcome of an random experiment
- Helps us define  $P(x)$

# Bernoulli Distribution (Discrete)

- Let  $x \in \{0, 1\}$  be the outcome of tossing a coin, we have:

$$\text{Bernoulli}(x = x; \rho) = \begin{cases} \rho, & \text{if } x = 1 \\ 1 - \rho, & \text{otherwise} \end{cases} \quad \text{or } \rho^x(1 - \rho)^{1-x}$$

- Properties: [Proof]
  - $E(x) = \rho$
  - $\text{Var}(x) = \rho(1 - \rho)$

# Categorical Distribution (Discrete)

- Let  $x \in \{1, \dots, k\}$  be the outcome of rolling a  $k$ -sided dice, we have:

$$\text{Categorical}(x = x; \rho) = \prod_{i=1}^k \rho_i^{1(x; x=i)}, \text{ where } \mathbf{1}^\top \rho = 1$$

# Categorical Distribution (Discrete)

- Let  $x \in \{1, \dots, k\}$  be the outcome of rolling a  $k$ -sided dice, we have:

$$\text{Categorical}(x = x; \rho) = \prod_{i=1}^k \rho_i^{1(x; x=i)}, \text{ where } \mathbf{1}^\top \rho = 1$$

- An extension of the Bernoulli distribution for  $k$  states



# Multinomial Distribution (Discrete)

- Let  $\mathbf{x} \in \mathbb{R}^k$  be a random vector where  $x_i$  the number of the outcome  $i$  after rolling a  $k$ -sided dice  $n$  times:

$$\text{Multinomial}(\mathbf{x} = \mathbf{x}; n, \boldsymbol{\rho}) = \frac{n!}{x_1! \cdots x_k!} \prod_{i=1}^k \rho_i^{x_i}, \text{ where } \mathbf{1}^\top \boldsymbol{\rho} = 1 \text{ and } \mathbf{1}^\top \mathbf{x} = n$$

# Multinomial Distribution (Discrete)

- Let  $\mathbf{x} \in \mathbb{R}^k$  be a random vector where  $x_i$  the number of the outcome  $i$  after rolling a  $k$ -sided dice  $n$  times:

$$\text{Multinomial}(\mathbf{x} = \mathbf{x}; n, \boldsymbol{\rho}) = \frac{n!}{x_1! \cdots x_k!} \prod_{i=1}^k \rho_i^{x_i}, \text{ where } \mathbf{1}^\top \boldsymbol{\rho} = 1 \text{ and } \mathbf{1}^\top \mathbf{x} = n$$

- Properties: [Proof]
  - $E(\mathbf{x}) = n\boldsymbol{\rho}$
  - $\text{Var}(\mathbf{x}) = n(\text{diag}(\boldsymbol{\rho}) - \boldsymbol{\rho}\boldsymbol{\rho}^\top)$   
(i.e.,  $\text{Var}(x_i) = n\rho_i(1 - \rho_i)$  and  $\text{Var}(x_i, x_j) = -n\rho_i\rho_j$ )

# Normal/Gaussian Distribution (Continuous)

## Theorem (Central Limit Theorem)

*The sum  $x$  of many independent random variables is approximately normally/Gaussian distributed:*

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

# Normal/Gaussian Distribution (Continuous)

## Theorem (Central Limit Theorem)

*The sum  $x$  of many independent random variables is approximately normally/Gaussian distributed:*

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

- Holds regardless of the original distributions of individual variables

# Normal/Gaussian Distribution (Continuous)

## Theorem (Central Limit Theorem)

*The sum  $x$  of many independent random variables is approximately normally/Gaussian distributed:*

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

- Holds regardless of the original distributions of individual variables
- $\mu_x = \mu$  and  $\sigma_x^2 = \sigma^2$

# Normal/Gaussian Distribution (Continuous)

## Theorem (Central Limit Theorem)

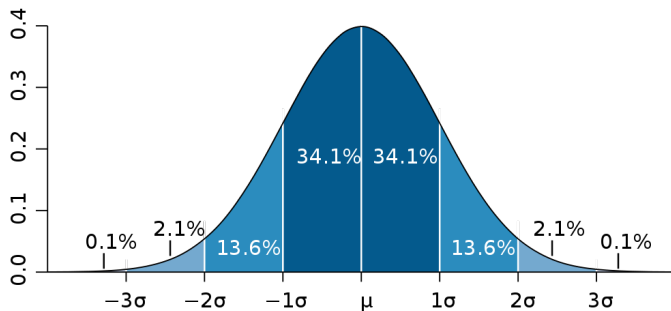
*The sum  $x$  of many independent random variables is approximately normally/Gaussian distributed:*

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

- Holds regardless of the original distributions of individual variables
- $\mu_x = \mu$  and  $\sigma_x^2 = \sigma^2$
- To avoid inverting  $\sigma^2$ , we can parametrize the distribution using the **precision**  $\beta$ :

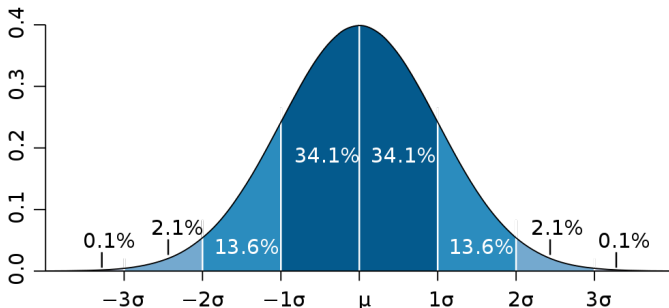
$$\mathcal{N}(x = x; \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(x - \mu)^2\right)$$

# Confidence Intervals



**Figure:** Graph of  $\mathcal{N}(\mu, \sigma^2)$ .

# Confidence Intervals



**Figure:** Graph of  $\mathcal{N}(\mu, \sigma^2)$ .

- We say the interval  $[\mu - 2\sigma, \mu + 2\sigma]$  has about the 95% confidence



# Why Is Gaussian Distribution So Common?

# Why Is Gaussian Distribution So Common?

- ① It can model complicate systems
  - E.g., Gaussian white noise

# Why Is Gaussian Distribution So Common?

- ① It can model complicate systems
  - E.g., Gaussian white noise
- ② Out of all possible probability distributions (over real numbers) with the same variance, it encodes the maximum amount of uncertainty
  - So, we insert the least amount of prior knowledge into a model

# Why Is Gaussian Distribution So Common?

- ① It can model complicate systems
  - E.g., Gaussian white noise
- ② Out of all possible probability distributions (over real numbers) with the same variance, it encodes the maximum amount of uncertainty
  - So, we insert the least amount of prior knowledge into a model
- ③ It is numerical friendly
  - E.g., continuous, differentiable, etc.

# Properties

- If  $x \sim \mathcal{N}(\mu, \sigma^2)$ , then  $ax + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$  for any deterministic  $a, b$  [Proof]
  - $z = \frac{x - \mu}{\sigma} \sim \mathcal{N}(0, 1)$  the **z-normalization** or **standardization** of  $x$

# Properties

- If  $x \sim \mathcal{N}(\mu, \sigma^2)$ , then  $ax + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$  for any deterministic  $a, b$  [Proof]
  - $z = \frac{x - \mu}{\sigma} \sim \mathcal{N}(0, 1)$  the **z-normalization** or **standardization** of  $x$
- If  $x^{(1)} \sim \mathcal{N}(\mu^{(1)}, \sigma^{2(1)})$  is independent with  $x^{(2)} \sim \mathcal{N}(\mu^{(2)}, \sigma^{2(2)})$ , then  $x^{(1)} + x^{(2)} \sim \mathcal{N}(\mu^{(1)} + \mu^{(2)}, \sigma^{2(1)} + \sigma^{2(2)})$   
[Homework:  $p_{x^{(1)}+x^{(2)}}(x) = \int p_{x^{(1)}}(x-y)p_{x^{(2)}}(y)dy$  the convolution]

# Properties

- If  $x \sim \mathcal{N}(\mu, \sigma^2)$ , then  $ax + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$  for any deterministic  $a, b$  [Proof]
  - $z = \frac{x - \mu}{\sigma} \sim \mathcal{N}(0, 1)$  the **z-normalization** or **standardization** of  $x$
- If  $x^{(1)} \sim \mathcal{N}(\mu^{(1)}, \sigma^{2(1)})$  is independent with  $x^{(2)} \sim \mathcal{N}(\mu^{(2)}, \sigma^{2(2)})$ , then  $x^{(1)} + x^{(2)} \sim \mathcal{N}(\mu^{(1)} + \mu^{(2)}, \sigma^{2(1)} + \sigma^{2(2)})$   
[Homework:  $p_{x^{(1)}+x^{(2)}}(x) = \int p_{x^{(1)}}(x-y)p_{x^{(2)}}(y)dy$  the convolution]
  - **Not** true if  $x^{(1)}$  and  $x^{(2)}$  are dependent

# Multivariate Gaussian Distribution

- When  $\mathbf{x}$  is sum of many random vectors:

$$\mathcal{N}(\mathbf{x} = \mathbf{x}; \mu, \Sigma) = \sqrt{\frac{1}{(2\pi)^d \det(\Sigma)}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

- $\mu_{\mathbf{x}} = \mu$  and  $\Sigma_{\mathbf{x}} = \Sigma$  (must be nonsingular)



# Multivariate Gaussian Distribution

- When  $\mathbf{x}$  is sum of many random vectors:

$$\mathcal{N}(\mathbf{x} = \mathbf{x}; \mu, \Sigma) = \sqrt{\frac{1}{(2\pi)^d \det(\Sigma)}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

- $\mu_{\mathbf{x}} = \mu$  and  $\Sigma_{\mathbf{x}} = \Sigma$  (must be nonsingular)
- If  $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ , then each attribute  $x_i$  is univariate normal
  - Converse **not** true

# Multivariate Gaussian Distribution

- When  $\mathbf{x}$  is sum of many random vectors:

$$\mathcal{N}(\mathbf{x} = \mathbf{x}; \mu, \Sigma) = \sqrt{\frac{1}{(2\pi)^d \det(\Sigma)}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

- $\mu_{\mathbf{x}} = \mu$  and  $\Sigma_{\mathbf{x}} = \Sigma$  (must be nonsingular)
- If  $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ , then each attribute  $x_i$  is univariate normal
  - Converse **not** true
  - However, if  $x_1, \dots, x_d$  are i.i.d. and  $x_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ , then  $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ , where  $\mu = [\mu_1, \dots, \mu_d]^\top$  and  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$

# Multivariate Gaussian Distribution

- When  $\mathbf{x}$  is sum of many random vectors:

$$\mathcal{N}(\mathbf{x} = \mathbf{x}; \mu, \Sigma) = \sqrt{\frac{1}{(2\pi)^d \det(\Sigma)}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

- $\mu_{\mathbf{x}} = \mu$  and  $\Sigma_{\mathbf{x}} = \Sigma$  (must be nonsingular)
- If  $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ , then each attribute  $x_i$  is univariate normal
  - Converse **not** true
  - However, if  $x_1, \dots, x_d$  are i.i.d. and  $x_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ , then  $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ , where  $\mu = [\mu_1, \dots, \mu_d]^\top$  and  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$
- What does the graph of  $\mathcal{N}(\mu, \Sigma)$  look like?

# Bivariate Example I

- Consider the *Mahalanobis distance* first

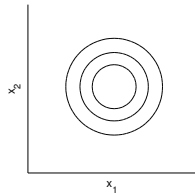
$$\mathcal{N}(\mu, \Sigma) = \sqrt{\frac{1}{(2\pi)^d \det(\Sigma)}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

# Bivariate Example I

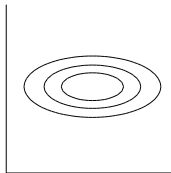
- Consider the *Mahalanobis distance* first

$$\mathcal{N}(\mu, \Sigma) = \sqrt{\frac{1}{(2\pi)^d \det(\Sigma)}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

$\text{Cov}(x_1, x_2) = 0, \text{Var}(x_1) = \text{Var}(x_2)$

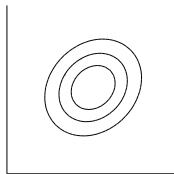


$\text{Cov}(x_1, x_2) = 0, \text{Var}(x_1) > \text{Var}(x_2)$

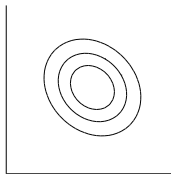


- The level sets closer to the center  $\mu_{\mathbf{x}}$  are lower

$\text{Cov}(x_1, x_2) > 0$



$\text{Cov}(x_1, x_2) < 0$

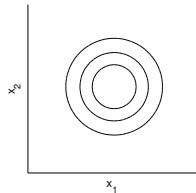


# Bivariate Example I

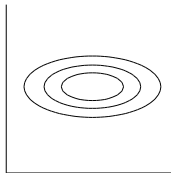
- Consider the *Mahalanobis distance* first

$$\mathcal{N}(\mu, \Sigma) = \sqrt{\frac{1}{(2\pi)^d \det(\Sigma)}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

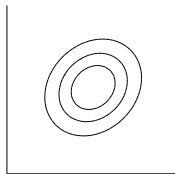
$\text{Cov}(x_1, x_2)=0, \text{Var}(x_1)=\text{Var}(x_2)$



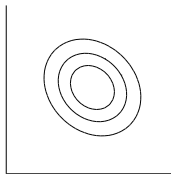
$\text{Cov}(x_1, x_2)=0, \text{Var}(x_1)>\text{Var}(x_2)$



$\text{Cov}(x_1, x_2)>0$



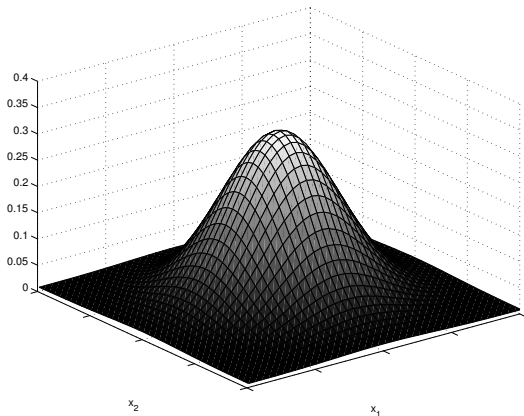
$\text{Cov}(x_1, x_2)<0$



- The level sets closer to the center  $\mu_{\mathbf{x}}$  are lower
- Increasing  $\text{Cov}[x_1, x_2]$  stretches the level sets along the  $45^\circ$  axis
- Decreasing  $\text{Cov}[x_1, x_2]$  stretches the level sets along the  $-45^\circ$  axis

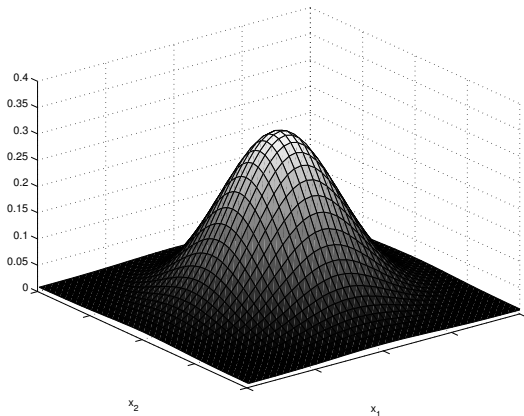
## Bivariate Example II

- The height of  $\mathcal{N}(\mu, \Sigma) = \sqrt{\frac{1}{(2\pi)^d \det(\Sigma)}} \exp \left[ -\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu) \right]$  in its graph is inversely proportional to the Mahalanobis distance



## Bivariate Example II

- The height of  $\mathcal{N}(\mu, \Sigma) = \sqrt{\frac{1}{(2\pi)^d \det(\Sigma)}} \exp \left[ -\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu) \right]$  in its graph is inversely proportional to the Mahalanobis distance



- A multivariate Gaussian distribution is *isotropic* iff  $\Sigma = \sigma I$



# Properties

- If  $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ , then  $\mathbf{w}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\top \mu, \mathbf{w}^\top \Sigma \mathbf{w})$  for any deterministic  $\mathbf{w} \in \mathbb{R}^d$

# Properties

- If  $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ , then  $\mathbf{w}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\top \mu, \mathbf{w}^\top \Sigma \mathbf{w})$  for any deterministic  $\mathbf{w} \in \mathbb{R}^d$
- More generally, given  $\mathbf{W} \in \mathbb{R}^{d \times k}$ ,  $k \leq d$ , we have  $\mathbf{W}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{W}^\top \mu, \mathbf{W}^\top \Sigma \mathbf{W})$  that is  $k$ -variate normal

# Properties

- If  $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ , then  $\mathbf{w}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\top \mu, \mathbf{w}^\top \Sigma \mathbf{w})$  for any deterministic  $\mathbf{w} \in \mathbb{R}^d$
- More generally, given  $\mathbf{W} \in \mathbb{R}^{d \times k}$ ,  $k \leq d$ , we have  $\mathbf{W}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{W}^\top \mu, \mathbf{W}^\top \Sigma \mathbf{W})$  that is  $k$ -variate normal
- The projection of  $\mathbf{x}$  onto a  $k$ -dimensional space is still normal

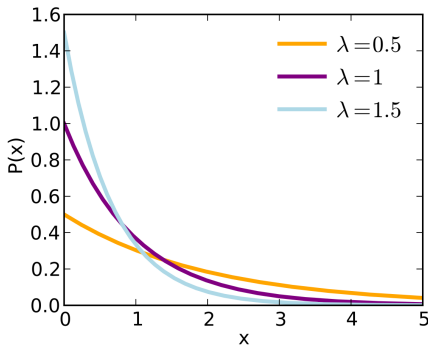
# Exponential Distribution (Continuous)

- In deep learning, we often want to have a probability distribution with a sharp point at  $x = 0$

# Exponential Distribution (Continuous)

- In deep learning, we often want to have a probability distribution with a sharp point at  $x = 0$
- To accomplish this, we can use the *exponential distribution*:

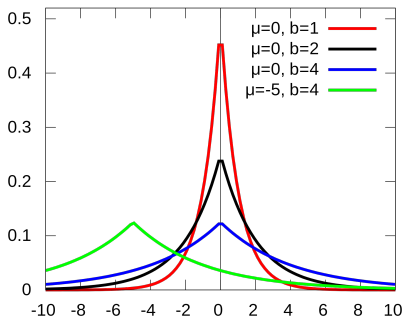
$$\text{Exponential}(x = x; \lambda) = \lambda 1(x; x \geq 0) \exp(-\lambda x)$$



# Laplace Distribution (Continuous)

- *Laplace distribution* can be think of as a “two-sided” exponential distribution centered at  $\mu$ :

$$\text{Laplace}(x = x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$



# Dirac Distribution (Continuous)

- In some cases, we wish to specify that all of the mass in a probability distribution clusters around a single data point  $\mu$

# Dirac Distribution (Continuous)

- In some cases, we wish to specify that all of the mass in a probability distribution clusters around a single data point  $\mu$
- This can be accomplished by using the *Dirac distribution*:

$$\text{Dirac}(\mathbf{x} = \mathbf{x}; \mu) = \delta(x - \mu),$$

where  $\delta(\cdot)$  is the Dirac delta function that

- ① Is zero-valued everywhere except at input  $\mathbf{0}$
- ② Integrals to 1



# Empirical Distribution (Continuous)

- Given a dataset  $\mathbb{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$  where  $\mathbf{x}^{(i)}$ 's are i.i.d. samples of  $\mathbf{x}$
- What is the distribution  $P(\theta)$  that maximizes the likelihood  $P(\theta|\mathbb{X})$  of  $\mathbb{X}$ ?

# Empirical Distribution (Continuous)

- Given a dataset  $\mathbb{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$  where  $\mathbf{x}^{(i)}$ 's are i.i.d. samples of  $\mathbf{x}$
- What is the distribution  $P(\theta)$  that maximizes the likelihood  $P(\theta|\mathbb{X})$  of  $\mathbb{X}$ ?
- If  $\mathbf{x}$  is discrete, the distribution simply reflects the empirical frequency of values:

$$\text{Empirical}(\mathbf{x} = \mathbf{x}; \mathbb{X}) = \frac{1}{N} \sum_{i=1}^N 1(\mathbf{x}; \mathbf{x} = \mathbf{x}^{(i)})$$

# Empirical Distribution (Continuous)

- Given a dataset  $\mathbb{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$  where  $\mathbf{x}^{(i)}$ 's are i.i.d. samples of  $\mathbf{x}$
- What is the distribution  $P(\theta)$  that maximizes the likelihood  $P(\theta|\mathbb{X})$  of  $\mathbb{X}$ ?
- If  $\mathbf{x}$  is discrete, the distribution simply reflects the empirical frequency of values:

$$\text{Empirical}(\mathbf{x} = \mathbf{x}; \mathbb{X}) = \frac{1}{N} \sum_{i=1}^N 1(\mathbf{x}; \mathbf{x} = \mathbf{x}^{(i)})$$

- If  $\mathbf{x}$  is continuous, we have the *empirical distribution*:

$$\text{Empirical}(\mathbf{x} = \mathbf{x}; \mathbb{X}) = \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}^{(i)})$$

# Mixtures of Distributions

- We may define a probability distribution by combining other simpler probability distributions  $\{P^{(i)}(\theta^{(i)})\}_i$

# Mixtures of Distributions

- We may define a probability distribution by combining other simpler probability distributions  $\{P^{(i)}(\theta^{(i)})\}_i$
- E.g., the *mixture model*:

$$\text{Mixture}(\mathbf{x} = \mathbf{x}; \rho, \{\theta^{(i)}\}_i) = \sum_i P^{(i)}(\mathbf{x} = \mathbf{x} | c = i; \theta^{(i)}) \text{Categorical}(c = i; \rho)$$

# Mixtures of Distributions

- We may define a probability distribution by combining other simpler probability distributions  $\{P^{(i)}(\theta^{(i)})\}_i$
- E.g., the *mixture model*:

$$\text{Mixture}(\mathbf{x} = \mathbf{x}; \rho, \{\theta^{(i)}\}_i) = \sum_i P^{(i)}(\mathbf{x} = \mathbf{x} | c = i; \theta^{(i)}) \text{Categorical}(c = i; \rho)$$

- The empirical distribution is a mixture distribution (where  $\rho_i = 1/N$ )

# Mixtures of Distributions

- We may define a probability distribution by combining other simpler probability distributions  $\{P^{(i)}(\theta^{(i)})\}_i$
- E.g., the *mixture model*:

$$\text{Mixture}(\mathbf{x} = \mathbf{x}; \rho, \{\theta^{(i)}\}_i) = \sum_i P^{(i)}(\mathbf{x} = \mathbf{x} | c = i; \theta^{(i)}) \text{Categorical}(c = i; \rho)$$

- The empirical distribution is a mixture distribution (where  $\rho_i = 1/N$ )
- The component identity variable  $c$  is a *latent variable*
  - Whose values are not observed

# Gaussian Mixture Model

- A mixture model is called the *Gaussian mixture model* iff
$$P^{(i)}(\mathbf{x} = \mathbf{x} | \mathbf{c} = i; \theta^{(i)}) = \mathcal{N}^{(i)}(\mathbf{x} = \mathbf{x} | \mathbf{c} = i; \mu^{(i)}, \Sigma^{(i)}), \forall i$$

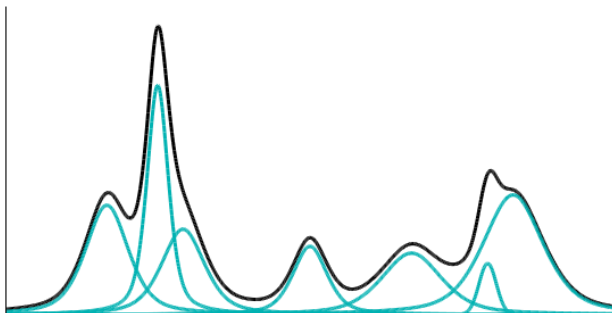


# Gaussian Mixture Model

- A mixture model is called the *Gaussian mixture model* iff
$$P^{(i)}(\mathbf{x} = \mathbf{x} | \mathbf{c} = i; \theta^{(i)}) = \mathcal{N}^{(i)}(\mathbf{x} = \mathbf{x} | \mathbf{c} = i; \mu^{(i)}, \Sigma^{(i)}), \forall i$$
  - Variants:  $\Sigma^{(i)} = \Sigma$  or  $\Sigma^{(i)} = \text{diag}(\sigma)$  or  $\Sigma^{(i)} = \sigma \mathbf{I}$

# Gaussian Mixture Model

- A mixture model is called the *Gaussian mixture model* iff
$$P^{(i)}(\mathbf{x} = \mathbf{x} | \mathbf{c} = i; \theta^{(i)}) = \mathcal{N}^{(i)}(\mathbf{x} = \mathbf{x} | \mathbf{c} = i; \mu^{(i)}, \Sigma^{(i)}), \forall i$$
  - Variants:  $\Sigma^{(i)} = \Sigma$  or  $\Sigma^{(i)} = \text{diag}(\sigma)$  or  $\Sigma^{(i)} = \sigma \mathbf{I}$
- Any smooth density can be approximated by a Gaussian mixture model with enough components



# Outline

- 1 Random Variables & Probability Distributions
- 2 Multivariate & Derived Random Variables
- 3 Bayes' Rule & Statistics
- 4 Application: Principal Components Analysis
- 5 Technical Details of Random Variables
- 6 Common Probability Distributions
- 7 Common Parametrizing Functions**
- 8 Information Theory
- 9 Application: Decision Trees & Random Forest

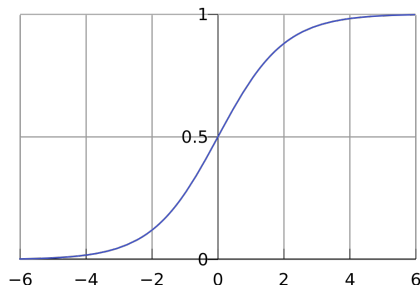
# Parametrizing Functions

- A probability distribution  $P(\theta)$  is parametrized by  $\theta$
- In ML,  $\theta$  may be the output value of a deterministic function
  - Called *parametrizing function*

# Logistic Function

- The *logistic function* (a special case of *sigmoid functions*) is defined as:

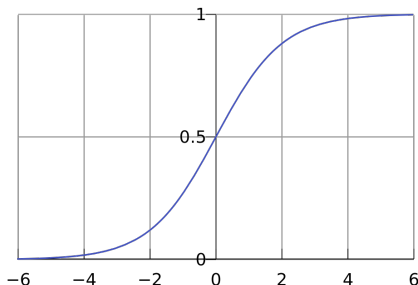
$$\sigma(x) = \frac{\exp(x)}{\exp(x) + 1} = \frac{1}{1 + \exp(-x)}$$



# Logistic Function

- The *logistic function* (a special case of *sigmoid functions*) is defined as:

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + 1} = \frac{1}{1 + \exp(-x)}$$

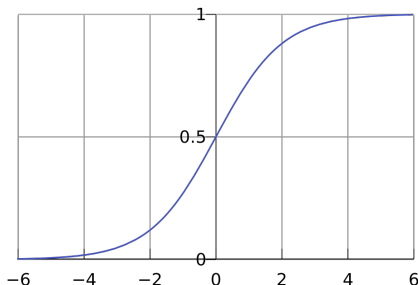


- Always takes on values between (0, 1)

# Logistic Function

- The *logistic function* (a special case of *sigmoid functions*) is defined as:

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + 1} = \frac{1}{1 + \exp(-x)}$$

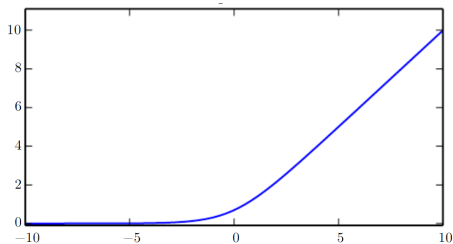


- Always takes on values between (0, 1)
- Commonly used to produce the  $\rho$  parameter of Bernoulli distribution

# Softplus Function

- The *softplus function* :

$$\zeta(x) = \log(1 + \exp(x))$$

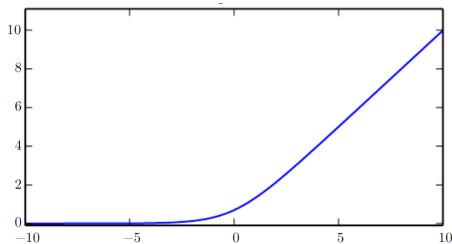




# Softplus Function

- The *softplus function* :

$$\zeta(x) = \log(1 + \exp(x))$$

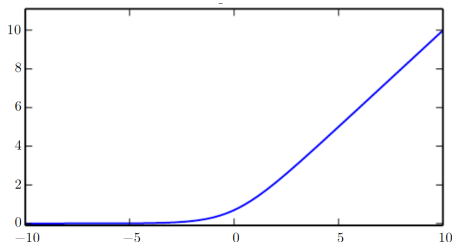


- A “softened” version of  $x^+ = \max(0, x)$

# Softplus Function

- The *softplus function* :

$$\zeta(x) = \log(1 + \exp(x))$$



- A “softened” version of  $x^+ = \max(0, x)$
- Range:  $(0, \infty)$
- Useful for producing the  $\beta$  or  $\sigma$  parameter of Gaussian distribution

# Properties [Homework]

- $1 - \sigma(x) = \sigma(-x)$
- $\log \sigma(x) = -\zeta(-x)$
- $\frac{d}{dx} \sigma(x) = \sigma(x)(1 - \sigma(x))$
- $\frac{d}{dx} \zeta(x) = \sigma(x)$
- $\forall x \in (0, 1), \sigma^{-1}(x) = \log\left(\frac{x}{1-x}\right)$
- $\forall x > 0, \zeta^{-1}(x) = \log(\exp(x) - 1)$
- $\zeta(x) = \int_{-\infty}^x \sigma(y) dy$
- $\zeta(x) - \zeta(-x) = x$ 
  - $\zeta(-x)$  is the softened  $x^- = \max(0, -x)$
  - $x = x^+ - x^-$

# Outline

- 1 Random Variables & Probability Distributions
- 2 Multivariate & Derived Random Variables
- 3 Bayes' Rule & Statistics
- 4 Application: Principal Components Analysis
- 5 Technical Details of Random Variables
- 6 Common Probability Distributions
- 7 Common Parametrizing Functions
- 8 Information Theory**
- 9 Application: Decision Trees & Random Forest

# What's Information Theory

- Probability theory allows us to make uncertain statements and reason in the presence of uncertainty

# What's Information Theory

- Probability theory allows us to make uncertain statements and reason in the presence of uncertainty
- Information theory allows us to *quantify* the amount of uncertainty

# Self-Information

- Given a random variable  $x$ , how much information you receive when seeing an event  $x = x$ ?

# Self-Information

- Given a random variable  $x$ , how much information you receive when seeing an event  $x = x$ ?
- ① Likely events should have low information
  - E.g., we are less surprised when tossing a biased coins



# Self-Information

- Given a random variable  $x$ , how much information you receive when seeing an event  $x = x$ ?
- ① Likely events should have low information
  - E.g., we are less surprised when tossing a biased coins
- ② Independent events should have additive information
  - E.g, “two heads” should have twice as much info as “one head”

# Self-Information

- Given a random variable  $x$ , how much information you receive when seeing an event  $x = x$ ?
- ① Likely events should have low information
  - E.g., we are less surprised when tossing a biased coins
- ② Independent events should have additive information
  - E.g, “two heads” should have twice as much info as “one head”
- The *self-information*:

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

# Self-Information

- Given a random variable  $x$ , how much information you receive when seeing an event  $x = x$ ?
- ① Likely events should have low information
  - E.g., we are less surprised when tossing a biased coins
- ② Independent events should have additive information
  - E.g., “two heads” should have twice as much info as “one head”
- The *self-information*:

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

- Called *bit* if base-2 logarithm is used
- Called *nat* if base- $e$

# Entropy

- Self-information deals with a particular outcome

# Entropy

- Self-information deals with a particular outcome
- We can quantify the amount of uncertainty in an entire probability distribution using the *entropy*:

$$H(\mathbf{x} \sim P) = E_{\mathbf{x} \sim P}[I(\mathbf{x})] = - \sum_x P(x) \log P(x) \text{ or } - \int p(x) \log p(x) dx$$

- Let  $0 \log 0 = \lim_{x \rightarrow 0} x \log x = 0$

# Entropy

- Self-information deals with a particular outcome
- We can quantify the amount of uncertainty in an entire probability distribution using the *entropy*:

$$H(x \sim P) = E_{x \sim P}[I(x)] = - \sum_x P(x) \log P(x) \text{ or } - \int p(x) \log p(x) dx$$

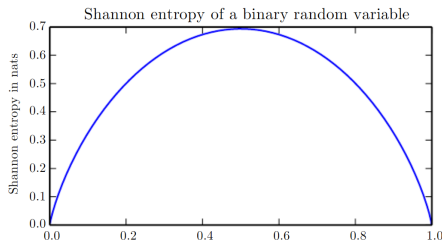
- Let  $0 \log 0 = \lim_{x \rightarrow 0} x \log x = 0$
- Called *Shannon entropy* when  $x$  is discrete; *differential entropy* when  $x$  is continuous

# Entropy

- Self-information deals with a particular outcome
- We can quantify the amount of uncertainty in an entire probability distribution using the **entropy**:

$$H(x \sim P) = E_{x \sim P}[I(x)] = - \sum_x P(x) \log P(x) \text{ or } - \int p(x) \log p(x) dx$$

- Let  $0 \log 0 = \lim_{x \rightarrow 0} x \log x = 0$
- Called **Shannon entropy** when  $x$  is discrete; **differential entropy** when  $x$  is continuous



**Figure:** Shannon entropy  $H(x)$  over Bernoulli distributions with different  $p$ .

# Average Code Length

- Shannon entropy gives a lower bound on the number of “bits” needed on average to encode values drawn from a distribution  $P$



# Average Code Length

- Shannon entropy gives a lower bound on the number of “bits” needed on average to encode values drawn from a distribution  $P$
- Consider a random variable  $x \sim \text{Uniform}$  having 8 equally likely states
  - To send a value  $x$  to receiver, we would encode it into 3 bits
  - Shannon entropy:  $H(x \sim \text{Uniform}) = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3$

# Average Code Length

- Shannon entropy gives a lower bound on the number of “bits” needed on average to encode values drawn from a distribution  $P$
- Consider a random variable  $x \sim \text{Uniform}$  having 8 equally likely states
  - To send a value  $x$  to receiver, we would encode it into 3 bits
  - Shannon entropy:  $H(x \sim \text{Uniform}) = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3$
- If the probabilities of the 8 states are  $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$  instead
  - $H(x) = 2$
  - The encoding 0, 10, 110, 1110, 111100, 111101, 111110, 111111 gives the average code length 2

# Kullback-Leibler (KL) Divergence

- How many extra “bits” needed in average to transmit a value drawn from distribution  $P$  when we use a code that was designed for another distribution  $Q$ ?

# Kullback-Leibler (KL) Divergence

- How many extra “bits” needed in average to transmit a value drawn from distribution  $P$  when we use a code that was designed for another distribution  $Q$ ?
- *Kullback-Leibler (KL) Divergence* or (*relative entropy*) from distribution  $Q$  to  $P$ :

$$D_{\text{KL}}(P\|Q) = E_{x \sim P} \left[ \log \frac{P(x)}{Q(x)} \right] = -E_{x \sim P} [\log Q(x)] - H(x \sim P)$$

- The term  $-E_{x \sim P} [\log Q(x)]$  is called the *cross entropy*

# Kullback-Leibler (KL) Divergence

- How many extra “bits” needed in average to transmit a value drawn from distribution  $P$  when we use a code that was designed for another distribution  $Q$ ?
- *Kullback-Leibler (KL) Divergence* or (*relative entropy*) from distribution  $Q$  to  $P$ :

$$D_{KL}(P\|Q) = E_{x \sim P} \left[ \log \frac{P(x)}{Q(x)} \right] = -E_{x \sim P} [\log Q(x)] - H(x \sim P)$$

- The term  $-E_{x \sim P} [\log Q(x)]$  is called the *cross entropy*
- If  $P$  and  $Q$  are independent, we can solve

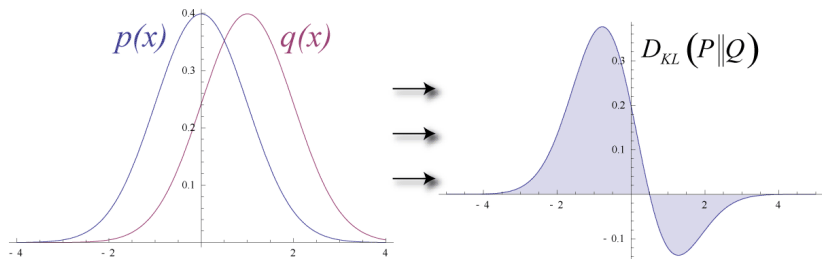
$$\arg \min_Q D_{KL}(P\|Q)$$

by

$$\arg \min_Q -E_{x \sim P} [\log Q(x)]$$

# Properties

- $D_{\text{KL}}(P\|Q) \geq 0, \forall P, Q$
- $D_{\text{KL}}(P\|Q) = 0$  iff  $P$  and  $Q$  are equal almost surely
- KL divergence is asymmetric, i.e.,  $D_{\text{KL}}(P\|Q) \neq D_{\text{KL}}(Q\|P)$



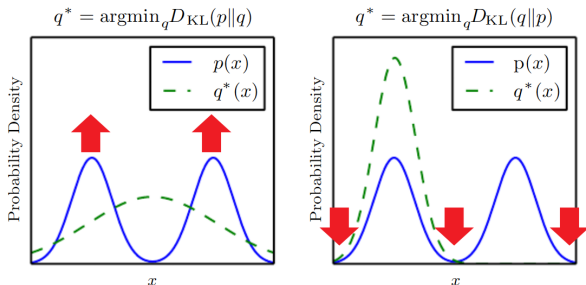
**Figure:** KL divergence for two normal distributions.

# Minimizer of KL Divergence

- Given  $P$ , we want to find  $Q^*$  that minimizes the KL divergence
- $Q^{*(\text{from})} = \arg \min_Q D_{\text{KL}}(P \| Q)$  or  $Q^{*(\text{to})} = \arg \min_Q D_{\text{KL}}(Q \| P)$ ?

# Minimizer of KL Divergence

- Given  $P$ , we want to find  $Q^*$  that minimizes the KL divergence
- $Q^{*(\text{from})} = \arg \min_Q D_{\text{KL}}(P \| Q)$  or  $Q^{*(\text{to})} = \arg \min_Q D_{\text{KL}}(Q \| P)$ ?
- $Q^{*(\text{from})}$  places high probability where  $P$  has high probability
- $Q^{*(\text{to})}$  places low probability where  $P$  has low probability



**Figure:** Approximating a mixture  $P$  of two Gaussians using a single Gaussian  $Q$ .

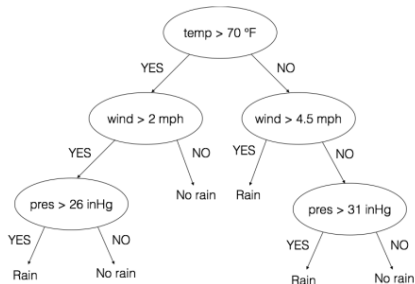


# Outline

- 1 Random Variables & Probability Distributions
- 2 Multivariate & Derived Random Variables
- 3 Bayes' Rule & Statistics
- 4 Application: Principal Components Analysis
- 5 Technical Details of Random Variables
- 6 Common Probability Distributions
- 7 Common Parametrizing Functions
- 8 Information Theory
- 9 Application: Decision Trees & Random Forest**

# Decision Trees

- Given a supervised dataset  $\mathbb{X} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$
- Can we find out a tree-like function  $f$  (i.e, a set of rules) such that  $f(\mathbf{x}^{(i)}) = y^{(i)}$ ?



# Training a Decision Tree

- Start from root which corresponds to all data points  
 $\{(\mathbf{x}^{(i)}, y^{(i)}) : \text{Rules} = \emptyset\}$
- Recursively split leaf nodes until data corresponding to children are “pure” in labels

# Training a Decision Tree

- Start from root which corresponds to all data points  
 $\{(\mathbf{x}^{(i)}, y^{(i)}) : \text{Rules} = \emptyset\}$
- Recursively split leaf nodes until data corresponding to children are “pure” in labels
- How to split?

# Training a Decision Tree

- Start from root which corresponds to all data points  
 $\{(\mathbf{x}^{(i)}, y^{(i)}) : \text{Rules} = \emptyset\}$
- Recursively split leaf nodes until data corresponding to children are “pure” in labels
- How to split? Find a cutting point  $(j, v)$  among all unseen attributes such that after partitioning the corresponding data points  
 $\mathbb{X}^{\text{parent}} = \{(\mathbf{x}^{(i)}, y^{(i)}) : \text{Rules}\}$  into two groups



$$\mathbb{X}^{\text{left}} = \{(\mathbf{x}^{(i)}, y^{(i)}) : \text{Rules} \cup \{x_j^{(i)} < v\}\}, \text{ and}$$

$$\mathbb{X}^{\text{right}} = \{(\mathbf{x}^{(i)}, y^{(i)}) : \text{Rules} \cup \{x_j^{(i)} \geq v\}\},$$

the “impurity” of labels drops the most

# Training a Decision Tree

- Start from root which corresponds to all data points  $\{(\mathbf{x}^{(i)}, y^{(i)}) : \text{Rules} = \emptyset\}$
- Recursively split leaf nodes until data corresponding to children are “pure” in labels
- How to split? Find a cutting point  $(j, v)$  among all unseen attributes such that after partitioning the corresponding data points  $\mathbb{X}^{\text{parent}} = \{(\mathbf{x}^{(i)}, y^{(i)}) : \text{Rules}\}$  into two groups



$$\mathbb{X}^{\text{left}} = \{(\mathbf{x}^{(i)}, y^{(i)}) : \text{Rules} \cup \{x_j^{(i)} < v\}\}, \text{ and}$$

$$\mathbb{X}^{\text{right}} = \{(\mathbf{x}^{(i)}, y^{(i)}) : \text{Rules} \cup \{x_j^{(i)} \geq v\}\},$$

the “impurity” of labels drops the most, i.e., solve

$$\arg \max_{j,v} \left( \text{Impurity}(\mathbb{X}^{\text{parent}}) - \text{Impurity}(\mathbb{X}^{\text{left}}, \mathbb{X}^{\text{right}}) \right)$$

# Impurity Measure

$$\arg \max_{j,v} \left( \text{Impurity}(\mathbb{X}^{\text{parent}}) - \text{Impurity}(\mathbb{X}^{\text{left}}, \mathbb{X}^{\text{right}}) \right)$$

- What's  $\text{Impurity}(\cdot)$ ?

# Impurity Measure

$$\arg \max_{j,v} \left( \text{Impurity}(\mathbb{X}^{\text{parent}}) - \text{Impurity}(\mathbb{X}^{\text{left}}, \mathbb{X}^{\text{right}}) \right)$$

- What's  $\text{Impurity}(\cdot)$ ?
- Entropy is a common choice:

$$\text{Impurity}(\mathbb{X}^{\text{parent}}) = H[y \sim \text{Empirical}(\mathbb{X}^{\text{parent}})]$$

$$\text{Impurity}(\mathbb{X}^{\text{left}}, \mathbb{X}^{\text{right}}) = \sum_{i=\text{left}, \text{right}} \frac{|\mathbb{X}^{(i)}|}{|\mathbb{X}^{\text{parent}}|} H[y \sim \text{Empirical}(\mathbb{X}^{(i)})]$$



# Impurity Measure

$$\arg \max_{j,v} \left( \text{Impurity}(\mathbb{X}^{\text{parent}}) - \text{Impurity}(\mathbb{X}^{\text{left}}, \mathbb{X}^{\text{right}}) \right)$$

- What's  $\text{Impurity}(\cdot)$ ?
- Entropy is a common choice:

$$\text{Impurity}(\mathbb{X}^{\text{parent}}) = H[y \sim \text{Empirical}(\mathbb{X}^{\text{parent}})]$$

$$\text{Impurity}(\mathbb{X}^{\text{left}}, \mathbb{X}^{\text{right}}) = \sum_{i=\text{left}, \text{right}} \frac{|\mathbb{X}^{(i)}|}{|\mathbb{X}^{\text{parent}}|} H[y \sim \text{Empirical}(\mathbb{X}^{(i)})]$$

- In this case,  $\text{Impurity}(\mathbb{X}^{\text{parent}}) - \text{Impurity}(\mathbb{X}^{\text{left}}, \mathbb{X}^{\text{right}})$  is called the *information gain*

# Random Forests

- A decision tree can be very deep

# Random Forests

- A decision tree can be very deep
- Deeper nodes give more specific rules
  - Backed by less training data
  - May not be applicable to testing data
- How to ensure the *generalizability* of a decision tree?
  - I.e., to have high prediction accuracy on testing data

# Random Forests

- A decision tree can be very deep
  - Deeper nodes give more specific rules
    - Backed by less training data
    - May not be applicable to testing data
  - How to ensure the *generalizability* of a decision tree?
    - I.e., to have high prediction accuracy on testing data
- ① Pruning (e.g., limit the depth of the tree)

# Random Forests

- A decision tree can be very deep
  - Deeper nodes give more specific rules
    - Backed by less training data
    - May not be applicable to testing data
  - How to ensure the *generalizability* of a decision tree?
    - I.e., to have high prediction accuracy on testing data
- ① Pruning (e.g., limit the depth of the tree)
  - ② *Random forest*: an ensemble of many (deep) trees

# Training a Random Forest

- ① Randomly pick  $M$  samples from the training set with replacement
  - Called the *bootstrap* samples

# Training a Random Forest

- ① Randomly pick  $M$  samples from the training set with replacement
  - Called the **bootstrap** samples
- ② Grow a decision tree from the bootstrap samples. At each node:
  - ① **Randomly select  $K$  features** without replacement
  - ② Find the best cutting point  $(j, v)$  and split the node

# Training a Random Forest

- ① Randomly pick  $M$  samples from the training set with replacement
  - Called the **bootstrap** samples
- ② Grow a decision tree from the bootstrap samples. At each node:
  - ① **Randomly select  $K$  features** without replacement
  - ② Find the best cutting point  $(j, v)$  and split the node
- ③ Repeat the steps 1 and 2 for  $T$  times to get  $T$  trees



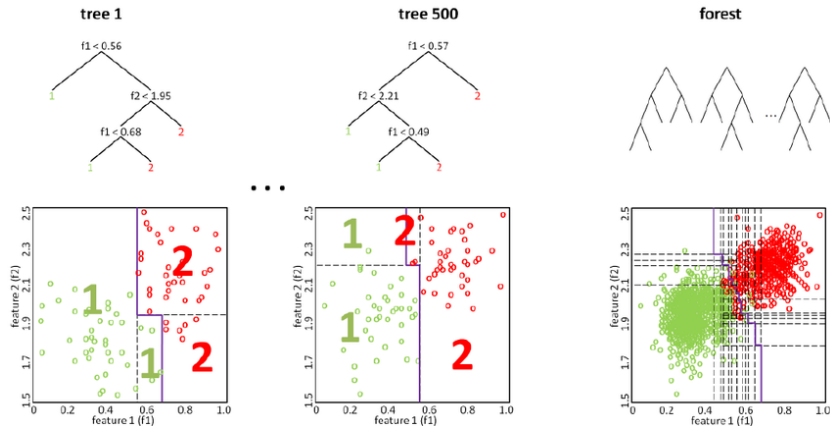
# Training a Random Forest

- ① Randomly pick  $M$  samples from the training set with replacement
  - Called the **bootstrap** samples
- ② Grow a decision tree from the bootstrap samples. At each node:
  - ① **Randomly select  $K$  features** without replacement
  - ② Find the best cutting point  $(j, v)$  and split the node
- ③ Repeat the steps 1 and 2 for  $T$  times to get  $T$  trees
- ④ Aggregate the predictions made by different trees via the **majority vote**

# Training a Random Forest

- ① Randomly pick  $M$  samples from the training set with replacement
  - Called the **bootstrap** samples
- ② Grow a decision tree from the bootstrap samples. At each node:
  - ① **Randomly select  $K$  features** without replacement
  - ② Find the best cutting point  $(j, v)$  and split the node
- ③ Repeat the steps 1 and 2 for  $T$  times to get  $T$  trees
- ④ Aggregate the predictions made by different trees via the **majority vote**
  - Each tree is trained slightly differently because of Step 1 and 2(a)
  - Provides different “perspectives” when voting

# Decision Boundaries



# Decision Trees vs. Random Forests

- Cons of random forests:
  - Less interpretable model

# Decision Trees vs. Random Forests

- Cons of random forests:
  - Less interpretable model
- Pros:
  - Less sensitive to the depth of trees
    - The majority voting can “absorb” the noise from individual trees
  - Can be parallelized
    - Each tree can grow independently