Probability & Information Theory

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

Department of Computer Science,
National Tsing Hua University, Taiwan

Machine Learning

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., $\Pr(\mathrm{x}=x_1)=0.1$, $\Pr(\mathrm{x}=x_2)=0.3$, 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
  - $\mathrm{x}\sim\mathrm{P}(\theta)$ means "x has distribution P parametrized by θ"

Probability Mass and Density Functions

- If x is discrete, $\mathrm{P}(\mathrm{x}=x)$ denotes a ***probability mass function*** $P_{\mathrm{x}}(x)=\Pr(\mathrm{x}=x)$
    - E.g., the output of a fair dice has discrete uniform distribution with $P(x)={}^{1}\!/_{6}$
- If x is continuous, $\mathrm{P}(\mathrm{x}=x)$ denotes a ***probability density function*** $p_{\mathrm{x}}(x)\geq 0$
    - Is $p_{\mathrm{x}}(x)$ a probability? ***No***, it is "rate of increase in probability at x"

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

-
    - $p_{\mathrm{x}}(x)$ 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., $\mathrm{P}(\mathrm{x},\mathrm{y})$
- The probability distribution over the subset of random variables called the ***marginal probability*** distribution:

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

  - Also called the sum rule of probability

Conditional Probability

- Conditional density function:

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

  - Defined only when $\mathrm{P}(\mathrm{y}=y)>0$

- Product rule of probability:

$$\mathrm{P}(\mathrm{x}^{(1)},\cdots,\mathrm{x}^{(n)})=\mathrm{P}(\mathrm{x}^{(1)})\Pi_{i=2}^{n}\mathrm{P}(\mathrm{x}^{(i)}\,|\,\mathrm{x}^{(1)},\cdots,\mathrm{x}^{(i-1)})$$

  - E.g., $\mathrm{P}(\mathrm{a},\mathrm{b},\mathrm{c})=\mathrm{P}(\mathrm{a}\,|\,\mathrm{b},\mathrm{c})\mathrm{P}(\mathrm{b}\,|\,\mathrm{c})\mathrm{P}(\mathrm{c})$

Independence and Conditional Independence

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

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

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

- We say random variables x is ***conditionally independent*** with y given z iff

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

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

Expectation

- The ***expectation*** (or ***expected value*** or ***mean***) of some function f with respect to x is the "average" value that f takes on:[1]

$$\mathrm{E}_{\mathrm{x}\sim\mathrm{P}}[\mathrm{f}(\mathrm{x})] = \sum_x P_{\mathrm{x}}(x) f(x) \text{ or } \int p_{\mathrm{x}}(x) f(x) dx = \mu_{\mathrm{f}(\mathrm{x})}$$

- Expectation is linear: $\mathrm{E}[a\mathrm{f}(\mathrm{x})+b] = a\mathrm{E}[\mathrm{f}(\mathrm{x})]+b$ for deterministic a and b
- $\mathrm{E}[\mathrm{E}[\mathrm{f}(\mathrm{x})]] = \mathrm{E}[\mathrm{f}(\mathrm{x})]$, as $\mathrm{E}[\mathrm{f}(\mathrm{x})]$ is deterministic

[1] 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.,

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

- $\mathrm{E}[\mathrm{f}(\mathrm{x}) \,|\, \mathrm{y} = y] = \int p_{\mathrm{x}|\mathrm{y}}(x \,|\, y) f(x) dx$ is called the ***conditional expectation***
- $\mathrm{E}[\mathrm{f}(\mathrm{x})\mathrm{g}(\mathrm{y})] = \mathrm{E}[\mathrm{f}(\mathrm{x})]\mathrm{E}[\mathrm{g}(\mathrm{y})]$ 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:

$$\mathrm{Var}[\mathrm{f}(\mathrm{x})] = \mathrm{E}\left[(\mathrm{f}(\mathrm{x}) - \mathrm{E}[\mathrm{f}(\mathrm{x})])^2\right] = \sigma^2_{\mathrm{f}(\mathrm{x})}$$

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

Covariance I

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

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

  - 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 $\mathrm{Cov}(\mathrm{x},\mathrm{y}) = 0$ [Proof]
  - The converse is ***not*** true as X and Y may be related in a nonlinear way
  - E.g., $\mathrm{y} = \sin(\mathrm{x})$ and $\mathrm{x} \sim \mathrm{Uniform}(-\pi,\pi)$

Covariance II

- $\mathrm{Var}(a\mathrm{x}+b\mathrm{y}) = a^2\mathrm{Var}(\mathrm{x}) + b^2\mathrm{Var}(\mathrm{y}) + 2ab\mathrm{Cov}(\mathrm{x},\mathrm{y})$ [Proof]
  - $\mathrm{Var}(\mathrm{x}+\mathrm{y}) = \mathrm{Var}(\mathrm{x}) + \mathrm{Var}(\mathrm{y})$ if x and y are independent
- $\mathrm{Cov}(a\mathrm{x}+b, c\mathrm{y}+d) = ac\mathrm{Cov}(\mathrm{x},\mathrm{y})$ [Proof]
- $\mathrm{Cov}(a\mathrm{x}+b\mathrm{y}, c\mathrm{w}+d\mathrm{v}) =$ $ac\mathrm{Cov}(\mathrm{x},\mathrm{w}) + ad\mathrm{Cov}(\mathrm{x},\mathrm{v}) + bc\mathrm{Cov}(\mathrm{y},\mathrm{w}) + bd\mathrm{Cov}(\mathrm{y},\mathrm{v})$ [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 $\mathbf{x} = [\mathrm{x}_1, \cdots, \mathrm{x}_d]^\top$
  - Normally, x_i's (***attributes*** or ***variables*** or ***features***) are dependent with each other
  - $\mathrm{P}(\mathbf{x})$ is a joint distribution of $\mathrm{x}_1, \cdots, \mathrm{x}_d$
- The ***mean*** of $\mathbf{x}$ is defined as $\mu_{\mathbf{x}} = \mathrm{E}(\mathbf{x}) = [\mu_{\mathrm{x}_1}, \cdots, \mu_{\mathrm{x}_d}]^\top$
- The ***covariance matrix*** of $\mathbf{x}$ is defined as:

$$\Sigma_{\mathbf{x}} = \begin{bmatrix} \sigma^2_{\mathrm{x}_1} & \sigma_{\mathrm{x}_1,\mathrm{x}_2} & \cdots & \sigma_{\mathrm{x}_1,\mathrm{x}_d} \\ \sigma_{\mathrm{x}_2,\mathrm{x}_1} & \sigma^2_{\mathrm{x}_2} & \cdots & \sigma_{\mathrm{x}_2,\mathrm{x}_d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\mathrm{x}_d,\mathrm{x}_1} & \sigma_{\mathrm{x}_d,\mathrm{x}_2} & \cdots & \sigma^2_{\mathrm{x}_d} \end{bmatrix}$$

  - $\sigma_{\mathrm{x}_i,\mathrm{x}_j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j) = \mathrm{E}[(\mathrm{x}_i - \mu_{\mathrm{x}_i})(\mathrm{x}_j - \mu_{\mathrm{x}_j})] = \mathrm{E}(\mathrm{x}_i\mathrm{x}_j) - \mu_{\mathrm{x}_i}\mu_{\mathrm{x}_j}$
  - $\Sigma_{\mathbf{x}} = \mathrm{Cov}(\mathbf{x}) = \mathrm{E}\left[(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})^\top\right] = \mathrm{E}(\mathbf{x}\mathbf{x}^\top) - \mu_{\mathbf{x}}\mu_{\mathbf{x}}^\top$

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 $\mathrm{y} = \mathrm{f}(\mathbf{x}; \boldsymbol{w}) = \boldsymbol{w}^\top \mathbf{x}$ be a random variable transformed from $\mathbf{x}$
- $\mu_\mathrm{y} = \mathrm{E}(\boldsymbol{w}^\top \mathbf{x}) = \boldsymbol{w}^\top \mathrm{E}(\mathbf{x}) = \boldsymbol{w}^\top \boldsymbol{\mu}_\mathbf{x}$
- $\sigma_\mathrm{y}^2 = \boldsymbol{w}^\top \Sigma_\mathbf{x} \boldsymbol{w}$ [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 $\Pr(\mathrm{x}=x)$ 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 $\Pr(\mathrm{x}=x)$

Bayes' Rule

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

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

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

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

Point Estimation

- ***Point estimation*** is the attempt to estimate some fixed but unknown quantity θ of a random variable by using sample data
- Let $\{x^{(1)},\cdots,x^{(n)}\}$ 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:

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

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

Sample Mean and Covariance

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

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

- A sample covariance matrix:

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

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

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

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

Principal Components Analysis (PCA) II

- Let $\boldsymbol{x}^{(i)}$'s be i.i.d. samples of a random variable $\mathbf{x}$
- Let f be linear, i.e., $f(\mathbf{x}) = \boldsymbol{W}^\top \mathbf{x}$ for some $\boldsymbol{W} \in \mathbb{R}^{D\times K}$
- ***Principal Component Analysis (PCA)*** finds K orthonormal vectors $\boldsymbol{W} = \left[\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(K)}\right]$ such that the transformed variable $\mathbf{z} = \boldsymbol{W}^\top \mathbf{x}$ has the most "spread out" attributes, i.e., each attribute $\mathrm{z}_j = \boldsymbol{w}^{(j)\top} \mathbf{x}$ has the maximum variance $\mathrm{Var}(\mathrm{z}_j)$
  - $\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(K)}$ are called the ***principle components***
- Why $\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(K)}$ need to be orthogonal with each other?
  - Each $\boldsymbol{w}^{(j)}$ keeps information that cannot be explained by others, so together they preserve the most info
- Why $\|\boldsymbol{w}^{(j)}\| = 1$ for all j?
  - Only directions matter—we don't want to maximize $\mathrm{Var}(\mathrm{z}_j)$ by finding a long $\boldsymbol{w}^{(j)}$

Solving W I

- For simplicity, let's consider $K = 1$ first
- How to evaluate $\mathrm{Var}(\mathrm{z}_1)$?
  - Recall that $\mathrm{z}_1 = \boldsymbol{w}^{(1)\top}\mathbf{x}$ implies $\sigma_{\mathrm{z}_1}^2 = \boldsymbol{w}^{(1)\top}\Sigma_{\mathbf{x}}\boldsymbol{w}^{(1)}$ [Homework]
  - How to get $\Sigma_{\mathbf{x}}$?
  - An estimate: $\hat{\Sigma}_{\mathbf{x}} = \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$ (assuming $\boldsymbol{x}^{(i)}$'s are centered first)
- Optimization problem to solve:

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

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

Solving $\boldsymbol{W}$ II

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

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

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

Visualization

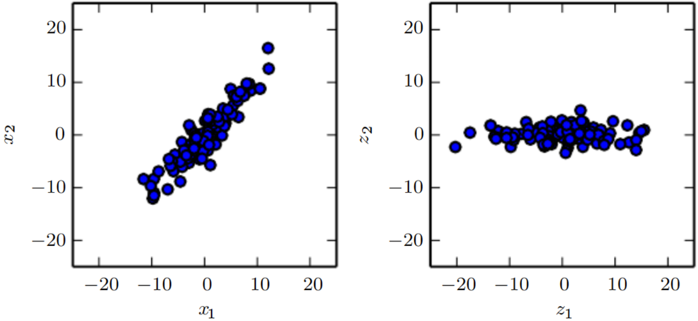


Figure: PCA learns a linear projection that aligns the direction of greatest variance with the axes of the new space. With these new axes, the estimated covariance matrix $\hat{\Sigma}_{\mathbf{z}} = \boldsymbol{W}^\top \hat{\Sigma}_{\mathbf{x}} \boldsymbol{W} \in \mathbb{R}^{K \times K}$ is always diagonal.

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 $\Pr(\mathrm{x}=x)=0$ for any value x
- Will the event $\mathrm{x}=x$ 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., $\Pr(\mathrm{x}\neq x)=1$)

Equality of Random Variables I

Definition (Equality in Distribution)

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

Definition (Almost Sure Equality)

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

Definition (Equality)

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

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 $P_{\mathrm{x}}(0) = P_{\mathrm{x}}(1) = P_{\mathrm{y}}(0) = P_{\mathrm{y}}(1) = 0.5$
  - They are equal in distribution
  - But $\Pr(\mathrm{x} = \mathrm{y}) = 0.5 \neq 1$

Convergence of Random Variables I

Definition (Convergence in Distribution)

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

Definition (Convergence in Probability)

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

Definition (Almost Sure Convergence)

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

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
- $\lim_{n\rightarrow\infty}\mathrm{Pr}\left(|\mathrm{x}^{(n)}-\mathrm{x}|<\varepsilon\right)=1$ leaves open the possibility that $|\mathrm{x}^{(n)}-\mathrm{x}|>\varepsilon$ happens an infinite number of times
- $\mathrm{Pr}\left(\lim_{n\rightarrow\infty}\mathrm{x}^{(n)}=\mathrm{x}\right)=1$ guarantees that $|\mathrm{x}^{(n)}-\mathrm{x}|>\varepsilon$ almost surely will not occur

Distribution of Derived Variables I

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

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

  - Violates the axiom of probability

Distribution of Derived Variables II

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

$$|p_{\mathrm{y}}(f(x))dy| = |p_{\mathrm{x}}(x)dx|$$

- We have

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

  - In previous example: $p_{\mathrm{y}}(y)=\mathbf{2}\cdot p_{\mathrm{x}}(2y)$

- In multivariate case, we have

$$p_{\mathbf{y}}(\boldsymbol{y})=p_{\mathbf{x}}(\boldsymbol{f}^{-1}(\boldsymbol{y}))\left|\det\left(\boldsymbol{J}(\boldsymbol{f}^{-1})(\boldsymbol{y})\right)\right|,$$

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

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

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 $\mathrm{P}(\mathrm{x})$

Bernoulli Distribution (Discrete)

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

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

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

Categorical Distribution (Discrete)

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

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

- 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:

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

- Properties: [Proof]
  - $\mathrm{E}(\mathbf{x}) = n\rho$
  - $\mathrm{Var}(\mathbf{x}) = n\left(\mathrm{diag}(\rho) - \rho\rho^\top\right)$
    $\big($i.e., $\mathrm{Var}(\mathrm{x}_i) = n\rho_i(1-\rho_i)$ and $\mathrm{Var}(\mathrm{x}_i,\mathrm{x}_j) = -n\rho_i\rho_j\big)$

Normal/Gaussian Distribution (Continuous)

Theorem (Central Limit Theorem)

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

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

- Holds regardless of the original distributions of individual variables
- $\mu_{\mathrm{x}}=\mu$ and $\sigma_{\mathrm{x}}^2=\sigma^2$
- To avoid inverting σ^2, we can parametrize the distribution using the ***precision*** β:

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

Confidence Intervals

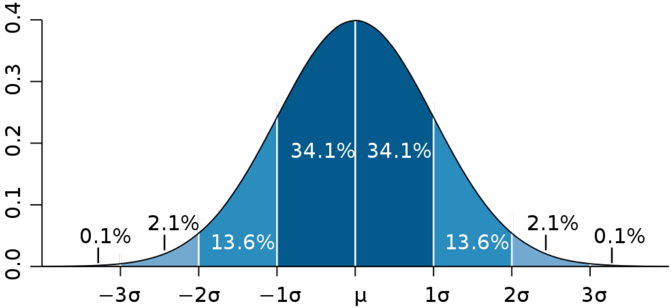


Figure: Graph of $\mathcal{N}(\mu,\sigma^2)$.

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

Why Is Gaussian Distribution Common in ML?

1. It can model noise in data (e.g., Gaussian white noise)
    - Can be considered to be the accumulation of a large number of small independent latent factors affecting data collection process
2. Out of all possible probability distributions (over real numbers) with the same variance, it encodes the maximum amount of uncertainty
    - Assuming $\mathrm{P}(\mathrm{y}\,|\,\mathrm{x}) \sim \mathscr{N}$, we insert the least amount of prior knowledge into a model
3. Convenient for many analytical manipulations
    - Closed under affine transformation, summation, marginalization, conditioning, etc.
    - Many of the integrals involving Gaussian distributions that arise in practice have simple closed form solutions

Properties

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

Multivariate Gaussian Distribution

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

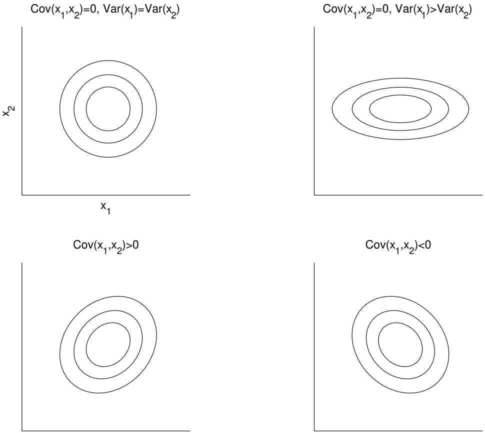
$$\mathcal{N}(\mathbf{x}=\boldsymbol{x};\boldsymbol{\mu},\Sigma)=\sqrt{\frac{1}{(2\pi)^d\det(\Sigma)}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

  - $\boldsymbol{\mu}_{\mathbf{x}}=\boldsymbol{\mu}$ and $\Sigma_{\mathbf{x}}=\Sigma$ (must be nonsingular)
- If $\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)$, then each attribute x_i is univariate normal
  - Converse ***not*** true
  - However, if $\mathrm{x}_1,\cdots,\mathrm{x}_d$ are i.i.d. and $\mathrm{x}_i\sim\mathcal{N}(\mu_i,\sigma_i^2)$, then $\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)$, where $\boldsymbol{\mu}=[\mu_1,\cdots,\mu_d]^\top$ and $\Sigma=\mathrm{diag}(\sigma_1^2,\cdots,\sigma_d^2)$
- What does the graph of $\mathcal{N}(\boldsymbol{\mu},\Sigma)$ look like?

Bivariate Example I

- Consider the ***Mahalanobis distance*** first

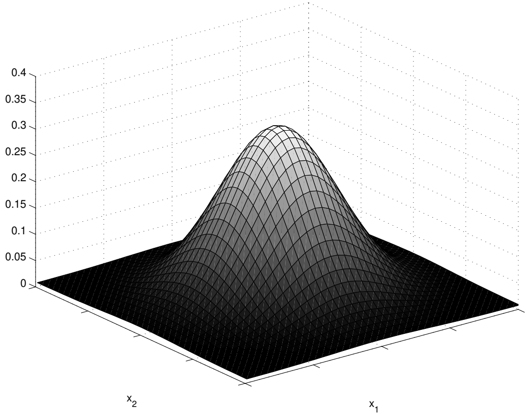
$$\mathcal{N}(\mu,\Sigma)=\sqrt{\frac{1}{(2\pi)^d\det(\Sigma)}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\mu)^\top\Sigma^{-1}(\boldsymbol{x}-\mu)\right]$$



- The level sets closer to the center $\mu_{\mathbf{x}}$ are lower
- Increasing $\mathrm{Cov}[\mathrm{x}_1,\mathrm{x}_2]$ stretches the level sets along the 45° axis
- Decreasing $\mathrm{Cov}[\mathrm{x}_1,\mathrm{x}_2]$ stretches the level sets along the -45° axis

Bivariate Example II

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



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

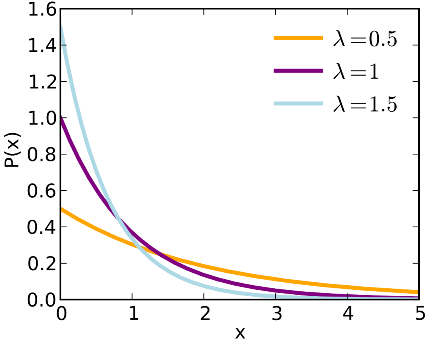
Properties

- Closed under affine transformation: if $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, then $\boldsymbol{w}^\top \mathbf{x} \sim \mathcal{N}(\boldsymbol{w}^\top \mu, \boldsymbol{w}^\top \Sigma \boldsymbol{w})$ for any deterministic $\boldsymbol{w} \in \mathbb{R}^d$
    - More generally, given $\boldsymbol{W} \in \mathbb{R}^{d \times k}$, $k < d$, we have $\boldsymbol{W}^\top \mathbf{x} \sim \mathcal{N}(\boldsymbol{W}^\top \mu, \boldsymbol{W}^\top \Sigma \boldsymbol{W})$ that is k-variate normal
    - I.e., the projection of $\mathbf{x}$ onto a k-dimensional subspace is still normal
- Consider $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim \mathcal{N}(\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{bmatrix})$:
- Closed under marginalization: $\mathbf{x}_1 \sim \mathcal{N}(\mu_1, \Sigma_{1,1})$ [Proof: $\mathrm{P}(\mathbf{x}_1) = \int_{\boldsymbol{x}_2} \mathrm{P}(\mathbf{x}_1, \boldsymbol{x}_2\,; \mu, \Sigma) d\boldsymbol{x}_2)$]
- Closed under conditioning: $(\mathbf{x}_1 \,|\, \boldsymbol{x}_2) \sim \mathcal{N}(\mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(\boldsymbol{x}_2 - \mu_2), \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})$ [Proof]

Exponential Distribution (Continuous)

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

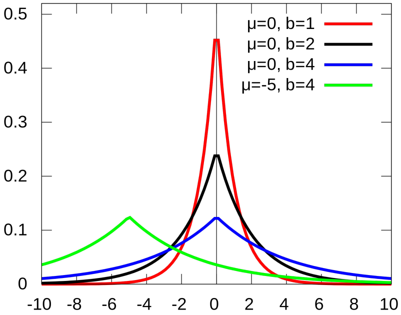
$$\text{Exponential}(\mathrm{x} = x; \lambda) = \lambda \mathbf{1}(x; x \geq 0) \exp(-\lambda x)$$



Laplace Distribution (Continuous)

- ***Laplace distribution*** can be think of as a "two-sided" exponential distribution centered at μ:

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



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 μ
- This can be accomplished by using the ***Dirac distribution***:

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

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

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

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

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

Mixtures of Distributions

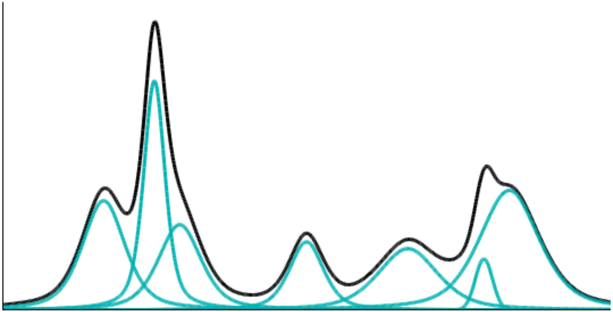
- We may define a probability distribution by combining other simpler probability distributions $\{\mathrm{P}^{(i)}(\theta^{(i)})\}_i$
- E.g., the ***mixture model***:

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

  - 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 $\mathrm{P}^{(i)}(\mathrm{x}=\boldsymbol{x}|\mathrm{c}=i;\boldsymbol{\theta}^{(i)})=\mathcal{N}^{(i)}(\mathrm{x}=\boldsymbol{x}|\mathrm{c}=i;\boldsymbol{\mu}^{(i)},\Sigma^{(i)})$, $\forall i$
  - Variants: $\Sigma^{(i)}=\Sigma$ or $\Sigma^{(i)}=\mathrm{diag}(\boldsymbol{\sigma})$ or $\Sigma^{(i)}=\boldsymbol{\sigma I}$
- 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

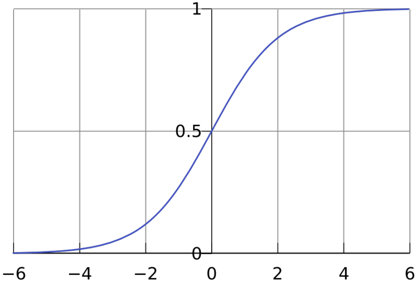
Parametrizing Functions

- A probability distribution $\mathrm{P}(\theta)$ is parametrized by θ
- In ML, θ 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:

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

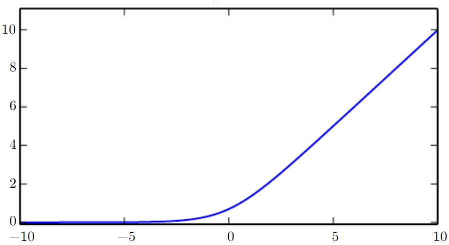


- Always takes on values between $(0,1)$
- Commonly used to produce the ρ parameter of Bernoulli distribution

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 β or σ 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
- 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 $\mathrm{x} = x$?

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***:

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

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

Entropy

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

$$\mathrm{H}(\mathrm{x} \sim \mathrm{P}) = \mathrm{E}_{\mathrm{x} \sim \mathrm{P}}[\mathrm{I}(\mathrm{x})] = -\sum_{x} P(x) \log P(x) \text{ or } -\int p(x) \log p(x) dx$$

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

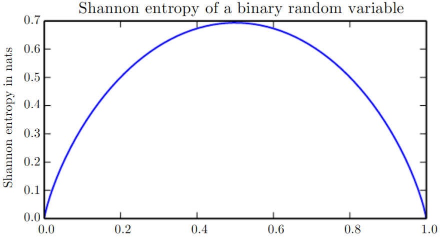


Figure: Shannon entropy $\mathrm{H}(\mathrm{x})$ over Bernoulli distributions with different ρ.

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 $\mathrm{x} \sim \mathrm{Uniform}$ having 8 equally likely states
  - To send a value x to receiver, we would encode it into 3 bits
  - Shannon entropy: $\mathrm{H}(\mathrm{x} \sim \mathrm{Uniform}) = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3$
- 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
  - $\mathrm{H}(\mathrm{x}) = 2$
  - 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*** or (***relative entropy***) from distribution Q to P:

$$\mathrm{D}_{\mathrm{KL}}(\mathrm{P}\|\mathrm{Q}) = \mathrm{E}_{\mathrm{x}\sim\mathrm{P}}\left[\log\frac{\mathrm{P}(\mathrm{x})}{\mathrm{Q}(\mathrm{x})}\right] = -\mathrm{E}_{\mathrm{x}\sim\mathrm{P}}[\log\mathrm{Q}(\mathrm{x})] - \mathrm{H}(\mathrm{x}\sim\mathrm{P})$$

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

$$\arg\min_{Q}\mathrm{D}_{\mathrm{KL}}(\mathrm{P}\|\mathrm{Q})$$

by

$$\arg\min_{Q}-\mathrm{E}_{\mathrm{x}\sim\mathrm{P}}[\log\mathrm{Q}(\mathrm{x})]$$

Properties

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

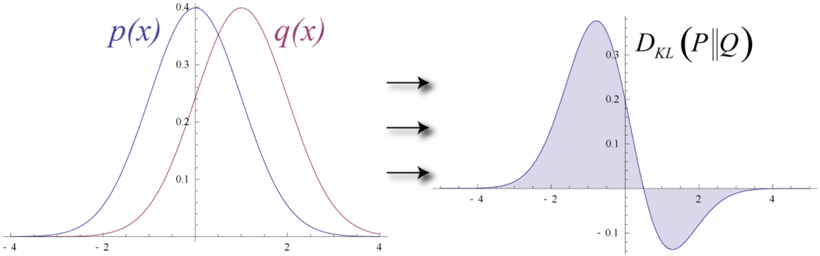


Figure: KL divergence for two normal distributions.

Minimizer of KL Divergence

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

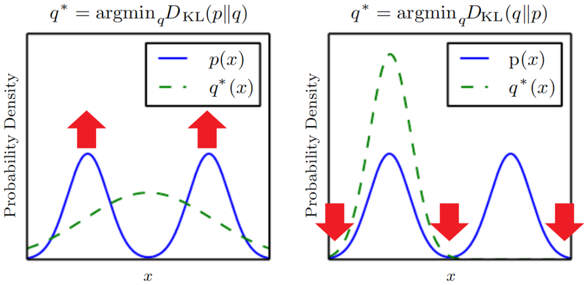


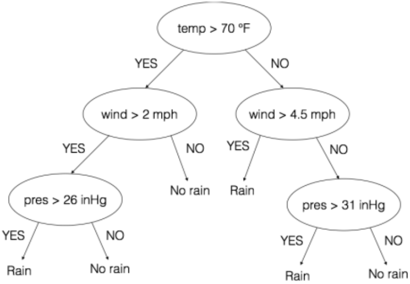
Figure: Approximating a mixture P of two Gaussians using a single Gaussian Q.

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

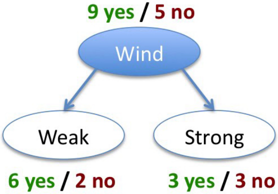


Training a Decision Tree

- Start from root which corresponds to all data points $\{(\boldsymbol{x}^{(i)}, y^{(i)}) : \text{Rules} = \emptyset)\}$
- 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 $\mathbb{X}^{\text{parent}} = \{(\boldsymbol{x}^{(i)}, y^{(i)} : \text{Rules})\}$ into two groups

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

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



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

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

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

- 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
- 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

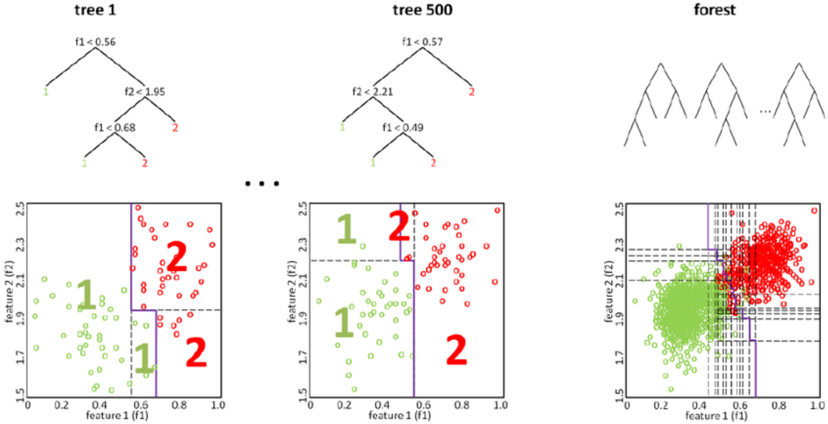
1. Pruning (e.g., limit the depth of the tree)
2. ***Random forest***: an ensemble of many (deep) trees

Training a Random Forest

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