Lecture 02: Univariate Probability

Introduction to Machine Learning [25737]

Sajjad Amini

Sharif University of Technology

Contents

References

The following reference inspires the material in the slides except cited:

- Murphy, K. P. (2022). *Probabilistic machine learning: an introduction.* MIT press.

Section 1

Probability Interpretations

Probability Interpretations

Frequentist (Long Run) Interpretation [1]

Probability are defined with respect to potentially infinite repetition of experiments. [2]

- Probability of heads in coin tossing: If we repeat the experiment of flipping a coin (at 'random'), the limit of the number of heads that occurred over the number of tosses is defined as the probability of a head occurring.'

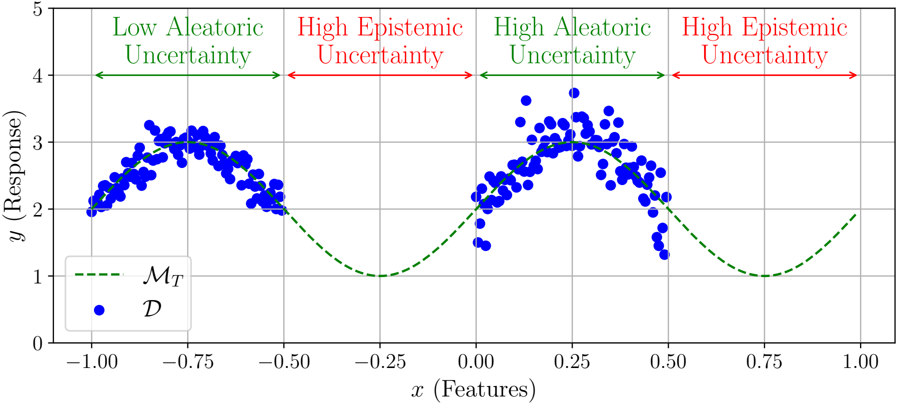
Probability Interpretations

Bayesian (Degree of Belief) Interpretation

Probability is a tool to quantify our uncertainty about something (This definition is fundamentally related to information rather than repeated trials.)

- The probability that a user likes or dislikes movies in the database:
  - This probability cannot be interpreted via repeated trials.
  - Assume that the user behaves consistently with other users. Then we can make a reasonable guess about whether he/she likes or dislikes the movie.

Uncertainty



Section 2

Random Variable

Random Variables and Events

Random Variable

Suppose X represents some quantity of interest. If the value of X is unknown and/or could change, we call it a Random Variable (RV).

Sample Space or State Space

The set of all possible values for Random variable X, denoted $\mathcal{X}$, is known as the sample space or state space.

Event

An event is a set of values from a random variable.

Examples

Random Variable

- X as the result of rolling a dice
- T as the room temperature

Sample Space or State Space

- $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ for random variable X
- $\mathcal{T} = \mathbb{R}$ for random variable T

Event

- Seeing an odd number in dice rolling experiment ($X \in \{1, 3, 5\}$)
- The temperature room is positive ($T \in \mathbb{R}^+$)

Discrete Random Variable

Discrete Random Variable

Random variable X is Discrete if its sample space $\mathcal{X}$ is finite or countably infinite.

Probability Mass Function

Consider x to be an arbitrary element in the sample space of random variable X. Probability mass function assigns $p(x)$ to x as:

$$p(x) \triangleq \Pr(X = x), \; x \in \mathcal{X}$$

Discrete Random Variable

Joint Distribution

Suppose a set of random variables $\{X_1, \ldots, X_n\}$. We can define the joint distribution of these random variables as:

$$p(x_1, \ldots, x_n) \triangleq \Pr(X_1 = x_1, \ldots, X_n = x_n), \quad \begin{cases} x_1 \in \mathcal{X}_1 \\ \vdots \\ x_n \in \mathcal{X}_n \end{cases}$$

Marginal Distribution

Given a joint distribution, we define the marginal distribution of random variable X_i as:

$$p(x_i) = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_{i-1} \in \mathcal{X}_{i-1}} \sum_{x_{i+1} \in \mathcal{X}_{i+1}} \cdots \sum_{x_n \in \mathcal{X}_n} p(x_1, \ldots, x_n)$$

Continuous Random Variable

Continuous Random Variable

Random variable X is Continuous if its sample space $\mathcal{X}$ is infinite and uncountable (Typically sample space is $\mathbb{R}$).

Cumulative Distribution Function

Consider x to be an arbitrary real value number. Cumulative Distribution Function assigns $P(x)$ to x as:

$$P(x) \triangleq \Pr(X \leq x),\ x \in \mathbb{R}$$

Probability Density Function (pdf)

Consider x to be an arbitrary real value number. Probability Density Function is defined using CDF as:

$$p(x) \triangleq \frac{d}{dx} P(x)$$

Continuous Random Variable

Joint Distribution

Suppose a set of random variables $\{X_1, \ldots, X_n\}$. We can define the joint distribution of these random variables as:

$$p(x_1, \ldots, x_n) \triangleq \frac{d^n}{dx_1 \ldots dx_n} \Pr(X_1 \leq x_1, \ldots, X_n \leq x_n), \quad \begin{cases} x_1 \in \mathbb{R} \\ \vdots \\ x_n \in \mathbb{R} \end{cases}$$

Marginal Distribution

Given a joint distribution, we define the marginal distribution of random variable X_i as:

$$p(x_i) = \int_{x_1=-\infty}^{\infty} \ldots \int_{x_{i-1}=-\infty}^{\infty} \int_{x_{i+1}=-\infty}^{\infty} \ldots \int_{x_n=-\infty}^{\infty} p(x_1, \ldots, x_n) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

Section 3

Bayes Rule

Conditional Probability (Bayes Rule)

Conditional Probability

The probability of event x conditioned on knowing event y is defined as:

$$p(x|y) \triangleq \frac{p(x,y)}{p(y)}$$

If $p(y) = 0$ then $p(x|y)$ is not defined. Equivalently we have:

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

Bayes Rule for Machine Learning

Naming Bayes Rule Factors

$$\overbrace{p(x|y)}^{\text{Posterior}} = \frac{\overbrace{p(y|x)}^{Likelihood} \quad \overbrace{p(x)}^{Prior}}{\underbrace{p(y)}_{Marginal}}$$

Bayes Rule for Machine Learning

Unsupervised learning

Replace $\begin{cases} x \rightarrow \boldsymbol{\theta} \\ y \rightarrow \{x_i\}_{i=1}^{N} \end{cases}$, then:

$$p(\boldsymbol{\theta}|\{x_i\}) = \frac{p(\{x_i\}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\{x_i\})}$$

Coin Tossing

Assume:

- X_i: Random variable representing the result of i-th tossing experiment
- θ: Bernoulli parameter

Then:

$$p(\theta|\{x_i\}) = \frac{p(\{x_i\}|\theta)p(\theta)}{p(\{x_i\})}$$

Bayes Rule for Machine Learning

Supervised learning

Replace $\begin{cases} x \rightarrow \boldsymbol{\theta} \\ y \rightarrow \{y_i\}_{i=1}^{N} \\ \text{Conditioning on } \{\boldsymbol{x}_i\}_{i=1}^{N} \end{cases}$, then:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\{y_i\},\{\boldsymbol{x}_i\}) &= \frac{p(\{y_i\}|\boldsymbol{\theta},\{\boldsymbol{x}_i\})p(\boldsymbol{\theta}|\{\boldsymbol{x}_i\})}{p(\{y_i\}|\{\boldsymbol{x}_i\})} \\
&= \frac{\left[\prod_i p(y_i|\boldsymbol{\theta},\boldsymbol{x}_i)\right] p(\boldsymbol{\theta})}{p(\{y_i\}|\{\boldsymbol{x}_i\})}
\end{aligned}
$$

Bayes Rule

Bayes Rule Interpreting [1]

Consider a dart board with 20 equal sections and the following RV:

$$X: \text{ Randy hit region 5}$$

- *Prior:* Randy hits any of 20 sections at random.
  - $p(X = 1) = \frac{1}{20}$
- *Knowledge (Evidence):* Randy hasn't hit region number 20.
- *Posterior:*

$$p(X = True|not\ 20) = \frac{p(X = True, not\ 20)}{p(not\ 20)} = \frac{p(X = True)}{p(not\ 20)}$$

$$= \frac{1/20}{19/20} = \frac{1}{19}$$

Section 4

Independence

Independence

Independence

Two random variable X and Y are unconditionally independent or marginally independent, denoted $X \perp Y$, iff we can represent the joint distribution as the product of the two marginal distribution. Thus we have:

$$X \perp Y \Leftrightarrow p(x, y) = p(x)p(y)$$

Equivalent Definitions

The following items are equivalent to independence:

- $p(x|y) = p(x)$
- $p(y|x) = p(y)$
- $p(x, y) = kf(x)g(y)$
  - k: constant
  - $f(\cdot)$: positive function
  - $g(\cdot)$: positive function

Independence

Independence [1]

Consider binary random variables X and Y with the following PMF:

$$p(X = a; Y = 1) = 1, \quad p(X = a; Y = 2) = 0$$
$$p(X = b; Y = 2) = 0; p(X = b; Y = 1) = 0$$

- $p(x)p(y) = p(x, y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, thus the RVs are independent.
- $\boldsymbol{X}$ and $\boldsymbol{Y}$ are always in the same joint state.

Independence

Independence [1]

Consider binary random variables X and Y with the following PMF:

$$p(X=a;Y=1)=1,\quad p(X=a;Y=2)=0$$
$$p(X=b;Y=2)=0; p(X=b;Y=1)=0$$

- $p(x)p(y)=p(x,y)$ for all $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, thus the RVs are independent.
- $\boldsymbol{X}$ and $\boldsymbol{Y}$ are always in the same joint state.

It is not a Contradiction

X and Y are independent if knowing the state of variable Y tells you something more than *you knew before* about variable X (*you knew before* means $p(x,y)$).

Conditional Independence

Conditional Independence

Two random variable X and Y are conditionally independent given Z, denoted $X \perp Y | Z$, if we can represent the conditional joint distribution as the product of the two conditional marginal distribution. Thus we have:

$$X \perp Y | Z \Leftrightarrow p(x, y|z) = p(x|z)p(y|z)$$

Empty Condition

If we have the following conditions:

- $X \perp Y | Z$
- $Z = \emptyset$

then X and Y are unconditionally independent.

Independence Implication

Independence Implication [1]

Suppose three random variables X, Y and Z. we have the following conditions:

- $X \perp Y$
- $Y \perp Z$

Does this conditions imply $X \perp Z$?

Independence Implication

Independence Implication [1]

Suppose three random variables X, Y and Z. we have the following conditions:

- $X \perp Y$
- $Y \perp Z$

Does this conditions imply $X \perp Z$?

Answer

NO! Assume $p(x, y, z) = p(y)p(x, z)$, then we can show clearly that the conditions hold while X is not necessarily independent of Z.

Conditional Independence [3]

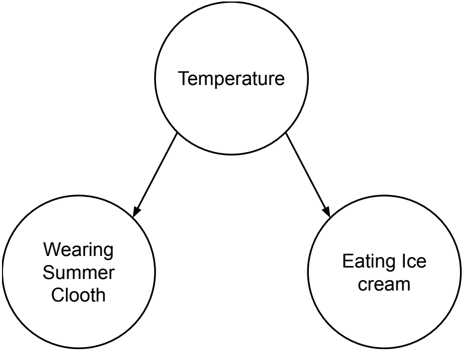


Figure: Conditional Independence (common cause)

Section 5

Probabilistic Reasoning

Probabilistic Reasoning (Inference) [1]

Probabilistic Reasoning

Consider the following two steps:

- Identifying all relevant random variables $X_1, \ldots, X_n$ in the environment
- Building a probabilistic model $p(x_1, \ldots, x_n)$ of their interaction

Then inference is performed by:

- Introducing *evidence* that sets some variables in known state
- Computing probabilities of interest, conditioned on the evidence.

Probabilistic Reasoning

The rules of probability combined with Bayes' rule make for a complete probabilistic reasoning system.

Hamburger [1]

Hamburger

Consider the following RVs:

- K: RV showing that a person have Kreuzfeld-Jacob disease (KJ)
- H: RV showing that a person is a hamburgers eater

We have also the following probabilities:

- *Prior:* $p(K=1) = \frac{1}{100000}$
- *Likelihood:* $p(H=1|K=1) = 0.9$

Suppose $p(H=1) = 0.5$. Whats is the probability of $p(K=1|H=1)$.
Solution:

$$p(K=1|H=1) = \frac{p(H=1, K=1)}{p(H=1)} = \frac{p(H=1|K=1)p(K=1)}{p(H=1)}$$

$$= \frac{\frac{9}{10} \times \frac{1}{100000}}{\frac{1}{2}} = 1.8 \times 10^{-5}$$

Hamburger [1]

Hamburger

Consider the following RVs:

- K: The probability that a person has Kreuzfeld-Jacob disease (KJ)
- H: The probability that a person is a hamburgers eater

We have also the following probabilities:

- *Prior:* $p(K=1) = \frac{1}{100000}$
- *Likelihood:* $p(H=1|K=1) = 0.9$

Suppose $p(H=1) = 0.001$. Whats is the probability of $p(K=1|H=1)$.

Solution:

$$p(K=1|H=1) = \frac{p(H=1, K=1)}{p(H=1)} = \frac{p(H=1|K=1)p(K=1)}{p(H=1)}$$

$$= \frac{\frac{9}{10} \times \frac{1}{100000}}{\frac{1}{1000}} \approx 1/100$$

Intuition: This example shows a stornger relation between eating hamburgers and KJ.

Inspector Challenge [1]

Inspector Challenge

Consider the following RVs:

- K: The murder uses a knife
- B: Butler is the murder
- M: Maid is the murder

Note that B and M are independent. We have also the following probabilities:

- *Prior* $p(B=1)=0.6,\ p(M=1)=0.2$
- *Likelihood*

$$p(K=1|B=0,M=0)=0.3,\quad p(K=1|B=0,M=1)=0.2$$
$$p(K=1|B=1,M=0)=0.6,\quad p(K=1|B=1,M=1)=0.1$$

Assume that the inspector finds that the murder was done using knife. What is the probability that Bob is the murder1.
Solution:

$$p(B=1|K=1)=\sum_m p(B=1,M=m|K=1)=\sum_m \frac{p(B=1,M=m,K=1)}{p(K=1)}$$
$$=\frac{p(B=1)\sum_m p(K=1|B=1,M=m)p(M=m)}{\sum_b p(B=b)\sum_m p(K=1|B=b,M=m)p(M=m)}\approx 0.73$$

Intuition: Knowing that the knife was the murder weapon strengthens our belief that the butler did it.

Resolution Reasoning

Resolution Reasoning

Resolution reasoning states that if $A \Rightarrow B$ and $B \Rightarrow C$, then we can infere $A \Rightarrow C$.

Resolution Reasoning

Consider the following statements:

- Statement A: *All apples are fruit* $\Rightarrow p(F=1|A=1)=1$
- Statement B: *All fruits grow on trees* $\Rightarrow p(T=1|F=1)=1$

Show that $p(T=1|A=1)=1$.

Solution:

$$p(T=1|A=1) = \sum_f p(T=1, F=f|A=1)) = \sum_f p(T=1|F=f, A=1)p(F=f|A=1)$$

$$= p(T=1|F=0)\overbrace{p(F=0|A=1)}^{=0} + \overbrace{p(T=1|F=1)}^{=1}\overbrace{p(F=1|A=1)}^{=1} = 1$$

Inverse Modus Ponens[1]

Inverse Modus Ponens

According to Logic, from the statement *If A is true then B is true*, one may deduce that *if B is false then A is false*.

Inverse Modus Ponens

Consider the following statement:

- *If A is true then B is true*: $p(B=1|A=1)=1$

Show that $p(A=0|B=0)=1$

Solution:

$$p(A=0|B=0) = 1 - p(A=1|B=0) = 1 - \frac{p(A=1,B=0)}{p(B=0)}$$
$$= 1 - \frac{p(B=0|A=1)p(A=1)}{p(B=0|A=1)p(A=1)+p(B=0|A=0)p(A=0)} = 1$$

Testing for COVID-19

COVID-19 Test Interpretation

Consider the following RVs:

- Y: RV showing that a person is infected with COVID-19
- X: RV showing person COVID-19 test result.

We have also the following probabilities:

- *Prior:* $p(Y=1)=0.1$ (prevalence of the disease in the area)
- *Likelihood:*

$$p(X=1|Y=1)=0.875,\ p(X=0|Y=0)=0.975$$

Calculate the posterior $p(Y=1|X=1)$ and $p(Y=1|X=0)$

Solution:

$$\begin{aligned} p(Y=1|X=1) &= \frac{p(X=1|Y=1)p(Y=1)}{p(X=1|Y=1)p(Y=1)+p(X=1|Y=0)p(Y=0)} \\ &= \frac{0.875\times 0.1}{0.875\times 0.1+0.025\times 0.9}=0.795 \\ p(Y=1|X=0) &= \frac{p(X=0|Y=1)p(Y=1)}{p(X=0|Y=1)p(Y=1)+p(X=0|Y=0)p(Y=0)} \\ &= \frac{0.125\times 0.1}{0.125\times 0.1+0.975\times 0.9}=0.014 \end{aligned}$$

Toward to Classification

Several Definitions:

We can assume the previous example as a binary classification problem where:

- Y : True state of infection
- X : Test result showing the state of infection

Based on this assumption we can have the following definitions:

- True Positive Rate (TPR) or Sensitivity: $p(X = 1|Y = 1)$
- True Negative Rate (TNR) or Specificity: $p(X = 0|Y = 0)$
- Flase Positive Rate (FPR): $p(X = 1|Y = 0) = 1 - TNR$
- Flase Negative Rate (FNR): $p(X = 0|Y = 1) = 1 - TPR$

Section 6

Sample PMF and Classification

Bernoulli Distribution

Bernoulli Distribution

Consider tossing a coin, where the probability of event that it lands heads is given by $0 \leq \theta \leq 1$. Let $Y = 1$ denote this event. Then random variable Y is distributed as Bernoulli distribution denoted by:

$$Y \sim Ber(\theta)$$

The PMF of this distribution is:

$$\begin{aligned} Ber(y|\theta) &= \begin{cases} 1-\theta & if\ y=0 \\ \theta & if\ y=1 \end{cases} \\ &= \theta^y (1-\theta)^{1-y} \end{aligned}$$

where $0 \leq \theta \leq 1$

Binomial Distribution

Binomial Distribution

Consider observing a set of N Bernoulli trials, denoted $Y_n \sim Ber(\cdot|\theta)$. Let us define random variable $S \triangleq \sum_{n=1}^{N} \mathbb{I}(Y_n = 1)$. Then random variable S is distributed as Binomial distribution denoted by:

$$Bin(s|N, \theta) \triangleq \binom{N}{s} \theta^s (1-\theta)^{N-s}$$

Bernoulli Distribution for Binary Classification

Classification Using Bernoulli Distribution

Suppose we want to predict a binary variable $y \in \{0, 1\}$ given some inputs $\boldsymbol{x} \in \mathcal{X}$. We can use Bernoulli Distribution to model conditional probability distribution as:

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = Ber(y|f(\boldsymbol{x}; \boldsymbol{\theta}))$$

where $0 \leq f(\boldsymbol{x}; \boldsymbol{\theta}) \leq 1$ is some function that predicts the mean parameter of the output distribution.

Sigmoid (Logistic) Function [4]

Sigmoid (Logistic) Function

Sigmoid (logistic) function, denoted $\sigma : \mathbb{R} \mapsto [0, 1]$, is defined as:$\sigma(a) \triangleq \frac{1}{1+e^{-a}}$

Sigmoid Function vs. Heaviside Step Function

The sigmoid function can be thought of as a *soft* version of the heaviside step function, defined by: $H(a) \triangleq \mathbb{I}(a > 0)$

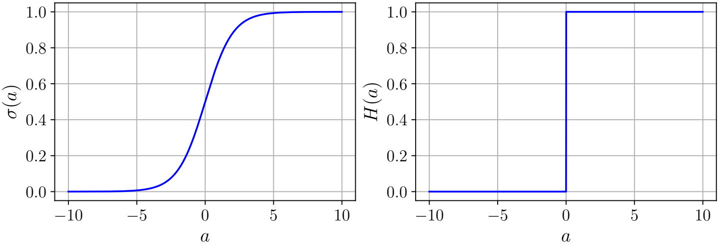


Figure: Sigmoid (Logistic) Function vs Heaviside Step Function

Bernoulli Distribution for Binary Classification

Classification Using Bernoulli Distribution

Suppose we want to predict a binary variable $y \in \{0, 1\}$ given some inputs $\boldsymbol{x} \in \mathcal{X}$. We can use Bernoulli Distribution to model conditional probability distribution as:

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = Ber(y|f(\boldsymbol{x}; \boldsymbol{\theta}))$$

To avoid $0 \leq f(\boldsymbol{x}; \boldsymbol{\theta}) \leq 1$ constraints, we can use the following conditional probability distribution:

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = Ber(y|\sigma(f(\boldsymbol{x}; \boldsymbol{\theta})))$$

Now $f(\boldsymbol{x}; \boldsymbol{\theta})$ is an arbitrary function.

Bernoulli Distribution for Binary Classification

From Sigmoid to Logit

Assume $a = f(\boldsymbol{x};\boldsymbol{\theta})$. Based on classification model $p(y|\boldsymbol{x},\boldsymbol{\theta}) = Ber(y|f(\boldsymbol{x};\boldsymbol{\theta}))$, we have:

$$p(y=1|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{1+e^{-a}} = \sigma(a)$$
$$p(y=0|\boldsymbol{x};\boldsymbol{\theta}) = 1 - \frac{1}{1+e^{-a}} = \sigma(-a)$$

Also if we define $p \triangleq p(y=1|\boldsymbol{x};\boldsymbol{\theta})$, we can calculate a as:

$$a = \sigma^{-1}(p) = \log\left(\frac{p}{1-p}\right)$$

Value a and function $\sigma^{-1}(\cdot)$ are known as *log odds* and *logit function*, respectively.

Iris Classification

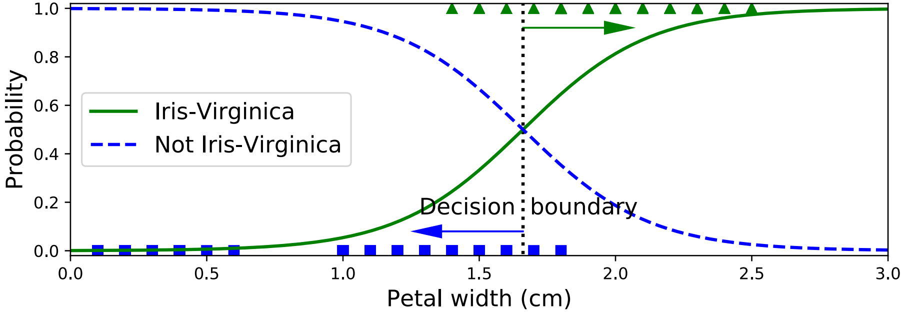


Figure: Iris classification using sigmoid function

Categorical Distribution

Categorical Distribution

Consider a distribution over a finite set of labels, $\mathcal{Y} = \{1, \ldots, C\}$. Let Y denote the label in one trial. Then random variable Y is distributed as Categorical distribution denoted by:

$$Y \sim Cat(\boldsymbol{\theta})$$

The PMF of this distribution is:

$$Cat(y|\boldsymbol{\theta}) \triangleq \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y=c)}$$

where $0 \leq \theta_c \leq 1$ and $\sum_{c=1}^{C} \theta_c = 1$.

Categorical Distribution Using one-hot Vector

One-hot encoding

$$\mathcal{Y} = \begin{cases} 1 & \rightarrow & [1, 0, 0, \ldots, 0, 0]^T \in \mathbb{R}^C \\ 2 & \rightarrow & [0, 1, 0, \ldots, 0, 0]^T \in \mathbb{R}^C \\ \vdots & & \vdots \\ C-1 & \rightarrow & [0, 0, 0, \ldots, 1, 0]^T \in \mathbb{R}^C \\ C & \rightarrow & [0, 0, 0, \ldots, 0, 1]^T \in \mathbb{R}^C \end{cases}$$

Categorical Distribution (Revisited)

If we define the one-hot coded vector $\boldsymbol{y}$ we have Categorical Distribution as:

$$Cat(\boldsymbol{y}|\boldsymbol{\theta}) \triangleq \prod_{c=1}^{C} \theta_c^{y_c}$$

where $0 \leq \theta_c \leq 1$ and $\sum_{c=1}^{C} \theta_c = 1$.

Multinomial Distribution

Consider observing a set of N Categorical trials, denoted $Y_n \sim Cat(\cdot|\boldsymbol{\theta})$. Let us define random vector $\boldsymbol{S} \triangleq \sum_{n=1}^{N} \boldsymbol{y}_n$. Then random vector $\boldsymbol{S}$ is distributed as Multinomial distribution denoted by:

$$Mu(\boldsymbol{s}|N,\boldsymbol{\theta}) \triangleq \binom{N}{s_1,\ldots,s_C} \prod_{c=1}^{C} \theta_c^{s_c}$$

where $\binom{N}{s_1,\ldots,s_C} \triangleq \frac{N!}{s_1!\ldots s_C!}$

Multinomial Distribution

For a multinomial distribution we have:

- $\sum_{c=1}^{C} s_c = N$
- If $N = 1$ then multinomial distribution becomes the categorical distribution.
- If $C = 2$ then multinomial distribution becomes the binomial distribution.

Categorical Distribution for Multiclass Classification

Classification Using Categorical Distribution

Suppose we want to predict one-hot coded vecotr of multiclass label $y \in \{1, \ldots, C\}$, denoted $\boldsymbol{y}$, given some inputs $\boldsymbol{x} \in \mathcal{X}$. We can use Categorical Distribution to model conditional probability distribution as:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = Cat(\boldsymbol{y}|\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}))$$

where $0 \leq f_c(\boldsymbol{x}; \boldsymbol{\theta}) \leq 1,\ c = 1, \ldots, C$ and $\sum_{c=1}^{C} f_c(\boldsymbol{x}; \boldsymbol{\theta}) = 1$. This function predicts the parameter of the output distribution.

Softmax Function [5]

Softmax Function

Softmax function, denoted $\boldsymbol{\mathcal{S}} : \mathbb{R}^C \mapsto [0, 1]^C$, is defined as:

$$\boldsymbol{\mathcal{S}}(\boldsymbol{a}) \triangleq \left[\frac{e^{a_1}}{\sum_{c'=1}^{C} e^{a_{c'}}}, \ldots, \frac{e^{a_C}}{\sum_{c'=1}^{C} e^{a_{c'}}} \right]$$

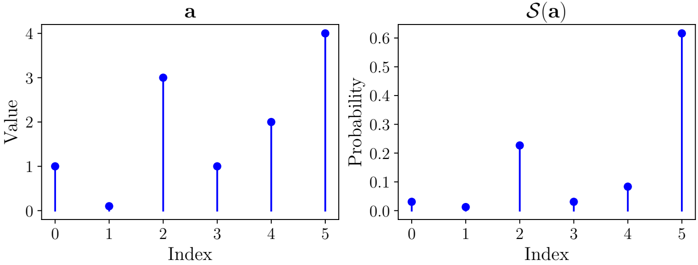


Figure: Softmax

Categorical Distribution for Multiclass Classification

Classification Using Categorical Distribution

Suppose we want to predict a variable $y \in \{1, \ldots, C\}$ given some inputs $\boldsymbol{x} \in \mathcal{X}$. We can use Categorical Distribution to model conditional probability distribution as:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = Cat(\boldsymbol{y}|\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}))$$

To avoid $0 \leq f_c(\boldsymbol{x}; \boldsymbol{\theta}) \leq 1, \; c = 1, \ldots, C$ and $\sum_{c=1}^{C} f_c(\boldsymbol{x}; \boldsymbol{\theta}) = 1$ constraints, we can use the following conditional probability distribution:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = Cat(\boldsymbol{y}|\boldsymbol{\mathcal{S}}(\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})))$$

Now $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})$ is an arbitrary function.

Section 7

Sample PDF

Gaussian Distribution [6]

Gaussian (Normal) Distribution

- The PDF for Gaussian (normal) distribution is:

$$\mathcal{N}(y|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

where μ and σ^2 are mean and variance, respectively.

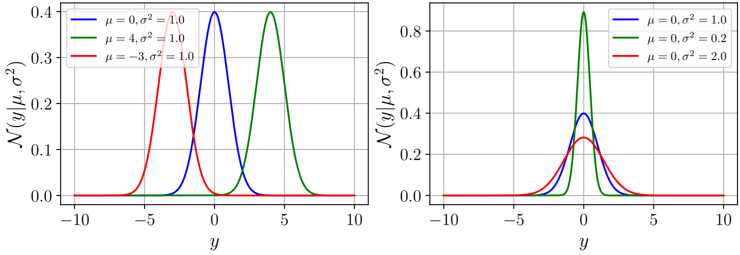


Figure: Normal Distribution

Laplace Distribution

- The PDF for Laplace distribution is:

$$\mathrm{Lap}(\mathrm{y}|\mu, \mathrm{b}) \triangleq \frac{1}{2\mathrm{b}} \mathrm{e}^{\left(-\frac{|\mathrm{y}-\mu|}{\mathrm{b}}\right)}$$

where μ and $b > 0$ are location and scale, respectively.

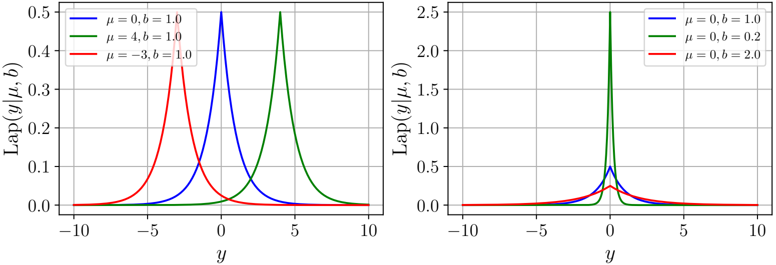


Figure: Laplace Distribution (Varying location and scale)

Section 8

Robust PDFs

Robust Distributions

Heavy-tailed distribution

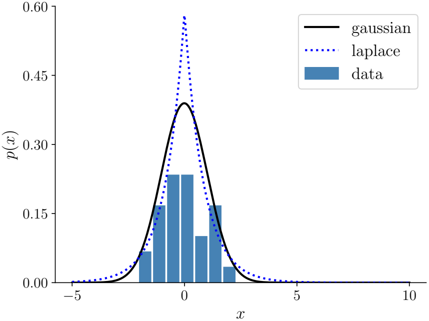
Assume random variable X. The right tail distribution function is defined as $\bar{f}(x) = Pr(X > x)$. Random variable X is sain to be right heavy tailed if:

$$\lim_{x \Rightarrow \infty} e^{tx} \bar{f}(x) = \infty$$

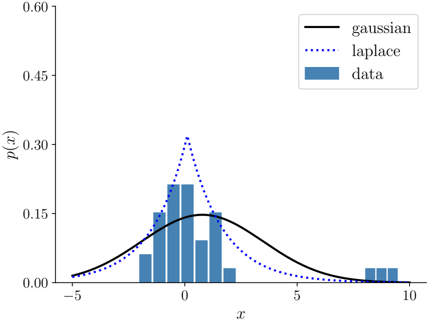
Heavy-tailed Distribution

Random variables with Student t and Laplace distributions are heavy-tailed while Gaussian random variable is light-tailed.

Robust Distributions



(a) Data without outlier



(b) Data with outlier

References I

David Barber,
Bayesian reasoning and machine learning,
Cambridge University Press, 2012.

“Epistemic uncertainty,” https://everyhue.me/media/3017/epistemic_uncertainty.png.

“Conditional independence,”
https://images.slideplayer.com/17/5382285/slides/slide_11.jpg.

“Sigmoid function,” https://miro.medium.com/max/875/1*ggtP4a5YaRx6l09KQaYOnw.png.

“Softmax function,” https://miro.medium.com/max/1400/1*ReYpdIZ3ZSAPb2W8cJpkBg.jpeg.

“Gaussian distribution,”
https://www.boost.org/doc/libs/1_38_0/libs/math/doc/sf_and_dist/graphs/normal_pdf.png.

“Laplace distribution,” https:
//www.statisticshowto.com/wp-content/uploads/2015/09/Laplace-distribution_pdf.png.