Probability &
Linear Algebra Review

Xuan Zhang
Sep 2, 2022
I. Probability
   • Random Variables
   • Bayes' Rule
     • Bernoulli distributions & Sigmoid function
     • Categorical distributions & Softmax function
     • Gaussian distribution

II. Bayes Optimal Rule

III. Linear Algebra
   • Vector norms
   • Matrix multiplication
   • Vector derivatives
I. Probability
Probability

Two interpretations:

- **Frequentist**
  probabilities represent long run frequencies of events

- **Bayesian (this course)**
  probability is used to quantify our uncertainty about something
  can be used to model uncertainty about one-off events
Probability

Event

\( \Pr (A) \): probability that \( A \) is true
\( \in [0, 1] \)

Joint probability: \( \Pr (A, B) \)

If \( A \) & \( B \) are independent, \( \Pr (A, B) = \Pr (A) \Pr (B) \)

Conditional probability: \( \Pr (B \mid A) \triangleq \frac{\Pr (A, B)}{\Pr (A)} \)

Conditionally independent \( A \perp B \mid C \)

\( \Pr (A, B \mid C) = \Pr (A \mid C) \Pr (B \mid C) \)
Random Variable (RV)

RVs are functions. RV is a numeric function of the outcome.

Sample space

HH → X(HH) = 2
HT → X(HT) = 1
TH → X(TH) = 1
TT → X(TT) = 0

X: number of heads
Random Variable (RV)

- **Discrete RV**: sample space is finite or countably infinite
- **Continuous RV**: infinite number of values between two values

**Cumulative distribution function (cdf)**

\[ P_c(x) \triangleq Pr(X \leq x) \]

e.g. \( Pr(a < X \leq b) = P_c(b) - P_c(a) \)

CDF's are monotonically non-decreasing functions
**Random Variable (RV)**

<table>
<thead>
<tr>
<th>Probability mass function (pmf) (discrete RV)</th>
<th>Probability density function (pdf) (continuous RV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(x) \triangleq \Pr(X = x) ) ( \sum_{x \in \mathcal{X}} p(x) = 1 )</td>
<td>( p(x) \triangleq \frac{d}{dx} p(x) ) ( \int_{-\infty}^{\infty} p(x) , dx = 1 )</td>
</tr>
</tbody>
</table>

**Examples:**
- **Uniform discrete distribution**
  - \( 
  \begin{array}{c}
  \text{1} \\
  \text{2} \\
  \text{3} \\
  \text{4} \\
  \end{array}
  \end{array}
  \)

**Graphs:**
- Uniform discrete distribution
- Gaussian pdf
Which of the following are valid cdg?

A. 

B. 

C. 

D. 

E. 

F.
Is it a CDF for a discrete random variable?
**Terminology**

- $X$: a random variable
- $x \sim X$: a sample value of a RV $X$
- $\Pr(x) = \Pr(X = x)$: probability of event $X$ has value $x$
- $\Pr(\bar{x}) = 1 - \Pr(x)$: probability of $x$ not happening
- $P(x)$: cumulative distribution function (cdf)
- $p(x)$: probability mass function (pmf)
- $f(x)$: probability density function (pdf)
Related Random Variables

Joint distribution:
\[ p(x, y) = p(X = x, Y = y) \]

Marginal distribution:
\[ p(X = x) = \sum_y p(X = x, Y = y) \rightarrow \text{sum rule} \]

Conditional distribution:
\[ p(Y = y | X = x) = \frac{p(X = x, Y = y)}{p(X = x)} \]
\[ p(x, y) = p(x) p(y | x) \rightarrow \text{product rule} \]

Chain rule of probability:
\[ p(X_1, 0) = p(x_1) \, p(x_2 | x_1) \, p(x_3 | x_1, x_2) \, \cdots \, p(x_0 | x_1, \ldots) \]
Moments of a distribution

- **Mean (expected values):**
  \[ E[X] = \sum_{x \in X} x \cdot p(x) \quad \text{or} \quad E[X] = \int x \cdot p(x) \, dx \]

  (discrete RV)  (continuous RV)

- **Linearity of expectation:**
  \[ E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] \]

- **Variance**
  \[ V[X] = E[(X - \mu)^2] = \int (x - \mu)^2 p(x) \, dx \]

  \[ = E[X^2] - \mu^2 \]

  \[ E[X^2] = \sigma^2 + \mu^2 \]

  \[ V[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} V[X_i] \]

  \[ V[aX + b] = a^2 V[X] \]

- **Mode**
  \[ x^* = \arg \max_x p(x) \]
Bayes' rule

$H$: unknown (or hidden) quantity

$Y = y$: observed data

prior: what we know about $H$ before seeing data

likelihood: distribution of data we expect to see if $H = h$

$p(H = h | Y = y) = \frac{p(H = h) \times p(Y = y | H = h)}{p(Y = y)}$

posterior: new belief state about $H$

posterior $\propto$ prior $\times$ likelihood
Example: The Monty Hall Problem

1. $1,000,000

2. ?

3. ?

$1,000,000

1. 2. 3.
Example: The Monty Hall Problem

What's your choice?
A. stick with door 1
B. switch to door 2
Example: The Monty Hall Problem

H: (hidden quantity) the prize is behind a door \{1, 2, 3\}
Y: observation a door is opened \{1, 2, 3\}

We want to compare \( P(H=1 | Y=3) \) vs. \( P(H=2 | Y=3) \)
Example: The Monty Hall Problem

- **H**: (hidden quantity) the prize is behind a door \(1, 2, 3\)
- **Y**: observation a door is opened \(1, 2, 3\)

We want to compare \(P(H=1 | Y=3)\) vs. \(P(H=2 | Y=3)\)

\[
P(H=1 | Y=3) = \frac{P(Y=3 | H=1) \cdot P(H=1)}{P(Y=3)} = \frac{1}{3}
\]

\[
P(Y=3) = P(Y=3 | H=1) \cdot P(H=1) + P(Y=3 | H=2) \cdot P(H=2) + P(Y=3 | H=3) \cdot P(H=3)
\]

\[
= \frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3} = \frac{1}{2}
\]

\[
P(H=2 | Y=3) = 1 - P(H=1 | Y=3) = \frac{2}{3}
\]

choose door 2!
## Bernoulli and Binomial Distribution

### Bernoulli distribution
- **toss a coin**
- $\theta$: probability that it lands head
  - $p(Y = 1) = \theta$
  - $p(Y = 0) = 1 - \theta$

  \[ Y \sim \text{Ber}(\theta) \]

**pmf:**

\[
\text{Ber}(y | \theta) = \begin{cases} 
1 - \theta & \text{if } y = 0 \\
\theta & \text{if } y = 1 
\end{cases}
\]

\[ \uparrow \]

### Binomial distribution
- **toss a coin $N$ times**

**number of heads:**

\[ S \triangleq \sum_{n=1}^{N} I(Y_n = 1) \]

\[ \text{Bin}(S | N, \theta) \triangleq \binom{N}{S} \theta^S (1-\theta)^{N-S} \]
Bernoulli Distribution and Sigmoid function

**Bernoulli distribution**

\[ \text{Ber}(y \mid \theta) = \theta^y (1 - \theta)^{1-y} \]

**Sigmoid (logistic) function**

We want to predict a binary variable \( y \in \{0, 1\} \) given inputs \( x \in \mathcal{X} \)

\[ p(y \mid x, \theta) = \text{Ber}(y \mid f(x; \theta)) \quad 0 \leq f(x; \theta) \leq 1 \]

The condition can be relaxed:

\[ p(y \mid x, \theta) = \text{Ber}(y \mid g(f(x; \theta))) \quad 0 \leq g(a) \leq 1 \]

where

\[ g(a) \triangleq \frac{1}{1 + e^{-a}} \]
**Sigmoid function**

- \( s(a) \triangleq \frac{1}{1 + e^{-a}} \), where \( a = \sigma(x; \theta) \)

- \( p(y=1 \mid x, \theta) = \frac{1}{1 + e^{-a}} = s(a) \)

- \( p(y=0 \mid x, \theta) = 1 - \frac{1}{1 + e^{-a}} = 1 - s(a) \)

**Log odds:**

\[
\log \left( \frac{p}{1-p} \right) = \log \left( \frac{e^a}{1 + e^a} \right) = \log(e^a) = a, \quad \text{where} \quad p = p(y=1 \mid x, \theta)
\]

**(Spoiler) Binary logistic regression**

\[
p(y \mid x, \theta) = \text{Ber}(y \mid s(w^T x + b)) \quad \Rightarrow \quad p(y=1 \mid x, \theta) = s(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}
\]

**Decision boundary:**

\( x^* \quad \Rightarrow \quad p(y=1 \mid x = x^*, \theta) = 0.5 \)
# Categorical and multinomial distributions

<table>
<thead>
<tr>
<th>Categorical Distribution</th>
<th>Multinomial Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>roll a $C$-sided dice</td>
<td>roll a $C$-sided dice $N$ times</td>
</tr>
<tr>
<td>$C &gt; 2$</td>
<td></td>
</tr>
</tbody>
</table>

**Categorical Distribution**

\[
\text{Cat}(y|\theta) \sim \frac{C}{\prod_{c=1}^{C} \theta_c^{\Pi(y=c)}}
\]

\[
p(y=c|\theta) = \theta_c
\]

**Multinomial Distribution**

\[
\text{Mn}(\leq N, \theta) \sim \binom{N}{s_1 \ldots s_c} \frac{C}{\prod_{c=1}^{C} \theta_c^{s_c}}
\]

where

\[
s_c = \sum_{n=1}^{N} \Pi(y_n = c)
\]

\[
\binom{N}{s_1 \ldots s_c} : \text{multinomial coefficient}
\]

number of ways to divide a set of size $N = \sum_{c=1}^{C} s_c$ into subsets with sizes $s_1$ to $s_c$
**Softmax function**

In conditional case:

\[
p(y|x, \theta) = \text{Cat}(y | f(x; \theta)) \quad 0 \leq f(x; \theta) \leq 1
\]

\[
p(y|x; \theta) = \text{Cat}(y | S(a)_c) \quad S(\cdot), \text{ softmax function} \quad 0 \leq S(a)_c \leq 1
\]

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

\[
a = f(x; \theta) \quad \text{logits}
\]
**Softmax function**

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

Temperature \( T \) : change the output distribution

as \( T \to 0 \), \( S(a/T)_c = \begin{cases} 1.0 & \text{if } c = \text{argmax}_c a_c, \\ 0.0 & \text{otherwise} \end{cases} \)
Gaussian (normal) distribution

\[ Y \sim N(\mu, \sigma^2) \]

pdf:

\[ N(y|\mu, \sigma^2) \propto \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \]

mean:

\[ \mu = \mathbb{E}[N(0.1|\mu, \sigma^2)] \]

standard deviation:

\[ \sigma = \text{std}[N(0.1|\mu, \sigma^2)] \]
II. Bayes Optimal Rule
Bayes Optimal Rule

We are searching for a \( f \) that minimizes the expected loss, measures distance between true label and prediction.

\[
\begin{align*}
\mathcal{J}^* &= \arg\min_f \mathbb{E}_{x,y} \left[ \text{loss}(y, f(x)) \right] \\
\text{prediction} \downarrow \\
\text{true label} \\
\text{population risk} \\
\text{can't compute because } P_{xy} \text{ is unknown.}
\end{align*}
\]
III. Linear Algebra
**Notation**

**vector** $x \in \mathbb{R}^n$

\[
x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}
\]

**one-hot vector (unit vector),**

$e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$

**matrix** $A \in \mathbb{R}^{m \times n}$

\[
A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}
\]

$A_{ij}$: the entry of $A$ in the $i$th row and $j$th column

$A_{i,:}$: the $i$th row

$A_{:,j}$: the $j$th column
Tensor

tensor = multidimensional array

vector \quad matrix \quad tensor

\textbf{v} \in \mathbb{R}^{64} \quad \textbf{X} \in \mathbb{R}^{8 \times 8} \quad \textbf{X} \in \mathbb{R}^{4 \times 4 \times 4}

tensor: a generalization of a 2d array to more than 2 dimensions.

order or rank of the tensor: the number of dimensions
**Vector Norms**

The norm of a vector $\|x\|$ is a measure of the "length" of the vector. The norm has the following properties:

1. **Non-negativity** $\|x\| \geq 0$
2. **Definiteness** $\|x\| = 0$ if $x = 0$
3. **Absolute value homogeneity** $\|tx\| = |t|\|x\|$ for all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$
4. **Triangle inequality** $\|x+y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$
Vector Norms

\( p \)-norm

\[ \| x \|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}, \text{ for } p \geq 1 \]

\( 2 \)-norm (Euclidean norm)

\[ \| x \|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \]

note that \( \| x \|_2^2 = x^T x \)

\( 1 \)-norm

\[ \| x \|_1 = \sum_{i=1}^{n} |x_i| \]

\( \infty \)-norm (max-norm)

\[ \| x \|_{\infty} = \max |x_i| \]
Matrix Multiplication

can only occur when \( A \in \mathbb{R}^{N \times M} \) \( B \in \mathbb{R}^{M \times D} \)

\[
A \times B = C \in \mathbb{R}^{N \times D}
\]

\[
C_{i,j} = \sum_{m=1}^{M} a_{i,m} \times b_{m,j}
\]

\[
C_{i,j} = A[i,:] \cdot B[:,j]
\]
# Matrix Derivatives and Gradients

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f = \left(\begin{array}{c} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{array}\right)$$

Consider a function $f : \mathbb{R}^{mn} \to \mathbb{R}$

$$\nabla f = \left(\begin{array}{c} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_{mn}} \end{array}\right)$$

Consider a function $f : \mathbb{R}^n \to \mathbb{R}^m$

$$J_f(x) = \nabla f = \left(\begin{array}{c} \frac{\partial f_1}{\partial x_1} \\ \vdots \\ \frac{\partial f_m}{\partial x_1} \\
\frac{\partial f_1}{\partial x_2} \\ \vdots \\ \frac{\partial f_m}{\partial x_2} \\
\vdots \\ \vdots \\ \frac{\partial f_1}{\partial x_n} \\ \vdots \\ \frac{\partial f_m}{\partial x_n} \end{array}\right) = \left(\begin{array}{c} \nabla f_1(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{array}\right)$$

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ that is twice differentiable

$$H_f = \frac{\partial^2 f}{\partial x^2} = \nabla^2 f = \left(\begin{array}{ccc} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{array}\right)$$
Matrix Derivatives  

\[
\frac{\partial (a^T x)}{\partial x} = a \quad \frac{\partial}{\partial x} (a^T x b) = ab^T \\
\frac{\partial (b^T A x)}{\partial x} = A^T b \quad \frac{\partial}{\partial x} (a^T x^T b) = ba^T \\
\frac{\partial (x^T A x)}{\partial x} = (A + A^T) x
\]

There are more!
Refer to <<The Matrix Cookbook>>, Petersen et al., Chapter 2.
Questions?