Probability Foundation (Continued) and Linear Algebra for Machine Learning

Dr. Xiaowei Huang

https://cgi.csc.liv.ac.uk/~xiaowei/

Demonstrators

- Jacopo Castellini
- Cameron Hargreaves
- Elektra Kypridemou
- Themistoklis Melissourgos

In the last week’s lectures,

- A few applications of machine learning
- Supervised vs. unsupervised learning
- Representation of instances as vectors
- Joint and conditional distribution

Topics of today

- Querying Joint Probability Distributions
  - Probability query
  - MAP query
- Scalars, vectors, matrices, tensors
- Multiplying matrices/vectors

Querying Joint Probability Distributions

Recap: Marginal, joint, conditional probability

- **Marginal probability**: the probability of an event occurring (p(A)), it may be thought of as an unconditional probability. It is not conditioned on another event.
  - Example: the probability that a card drawn is red (p(red) = 0.5).
  - Another example: the probability that a card drawn is a 4 (p(four)=1/13).

- **Joint probability**: p(A and B). The probability of event A **and** event B occurring. It is the probability of the intersection of two or more events. The probability of the intersection of A and B may be written $p(A \cap B)$.
  - Example: the probability that a card is a four and red =p(four and red) = 2/52=1/26. (There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).

- **Conditional probability**: $p(A|B)$ is the probability of event A occurring, given that event B occurs.
  - Example: given that you drew a red card, what's the probability that it's a four (p(four|red))=2/26=1/13. So out of the 26 red cards (given a red card), there are two fours so 2/26=1/13.

Recap: Chain Rules

chain rule (also called the **general product rule**[1][2]) permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities.

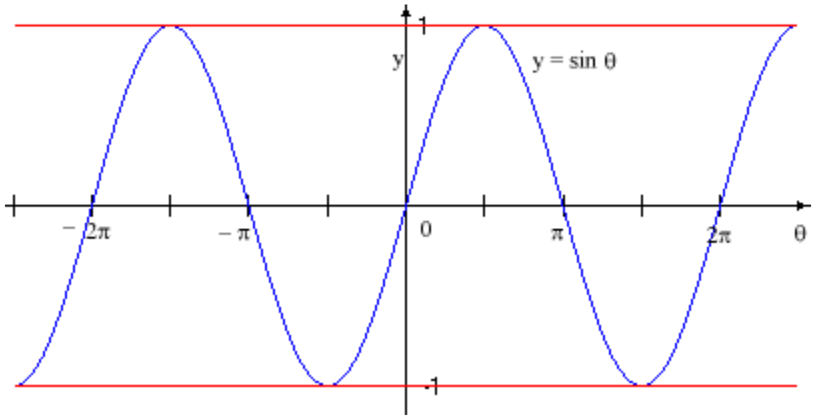
$$\mathrm{P}(A_n, \ldots, A_1) = \mathrm{P}(A_n | A_{n-1}, \ldots, A_1) \cdot \mathrm{P}(A_{n-1}, \ldots, A_1)$$

$$\mathrm{P}(A_4, A_3, A_2, A_1) = \mathrm{P}(A_4 \mid A_3, A_2, A_1) \cdot \mathrm{P}(A_3 \mid A_2, A_1) \cdot \mathrm{P}(A_2 \mid A_1) \cdot \mathrm{P}(A_1)$$

Recap: Max vs. argmax

- Let x be in a range [a,b] and f be a function over [a,b], we have
  - max f(x) to represent the maximum value of f(x) as x varies through [a,b]
  - argmax f(x) to represent the value of x at which the maximum is attained

- $\max_x \sin(x)$
  $= 1$
- $\text{argmax}_x \sin(x)$
  = {(0.5+2n)*pi | n is integer }
  = {…, -1.5pi, 0.5pi, 2.5pi, …}



Query Types

- Probability Queries
  - Given evidence (the values of a subset of random variables),
  - compute distribution of another subset of random variables
- MAP Queries
  - Maximum a posteriori probability
  - Also called MPE (*Most Probable Explanation*)
    - What is the most likely setting of a subset of random variables
  - Marginal MAP Queries
    - When some variables are known

Probability Queries

- Most common type of query is a probability query
- Query has two parts
  - *Evidence*: a subset *E* of variables and their instantiation *e*
  - *Query Variables*: a subset *Y* of random variables
- Inference Task: P(Y|E=e)
  - *Posterior probability distribution* over values *y* of *Y*
  - *Conditioned* on the fact *E=e*
  - Can be viewed as Marginal over *Y* in distribution we obtain by conditioning on *e*
- Marginal Probability Estimation $P(Y = y_i \mid E = e) = \dfrac{P(Y = y_i, E = e)}{P(E = e)}$

MAP Queries (Most Probable Explanation)

- Finding a high probability assignment to some subset of variables
- Most likely assignment to all non-evidence variables *W* = V – *E*

$$MAP(W \mid e) = \arg\max_{w} P(w, e)$$

i.e., value of *w* for which *P(w,e)* is maximum

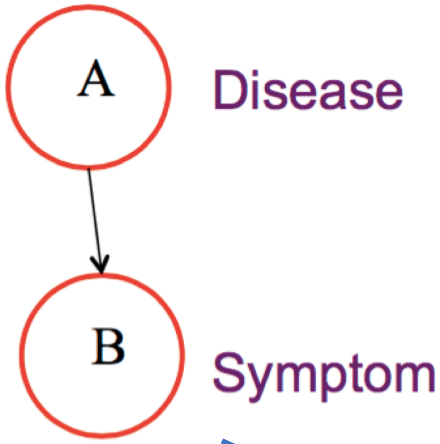
- Difference from probability query
  - Instead of a probability we get the most likely value for all remaining variables

Example of MAP Queries

- Medical Diagnosis Problem
  - Diseases (A) cause Symptoms (B)
  - Two possible diseases: Mono and Flu
  - Two possible symptoms: Headache and Fever

P(Diseases)

a^0	a^1
0,4	0.6



P(Symptom|Disease)

$P(B\|A)$	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Notation for probabilistic graphical models, to be introduced in later part of this module

Example of MAP Queries

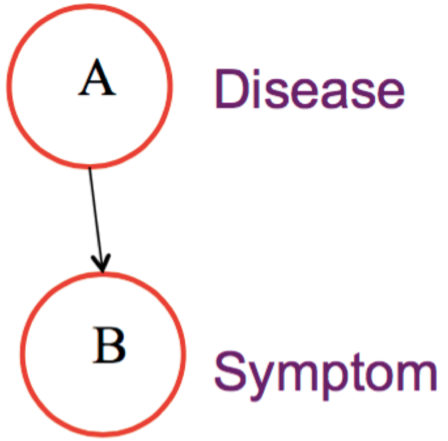
- Medical Diagnosis Problem
  - Diseases (*A*) cause Symptoms (*B*)
  - Two possible diseases: Mono and Flu
  - Two possible symptoms: Headache and Fever

- Q1: Most likely disease *P(A)*?

$$\mathrm{MAP}(\mathrm{A}) = \arg\max_a A \quad = a^1$$

P(Diseases)

a^0	a^1
0,4	0.6



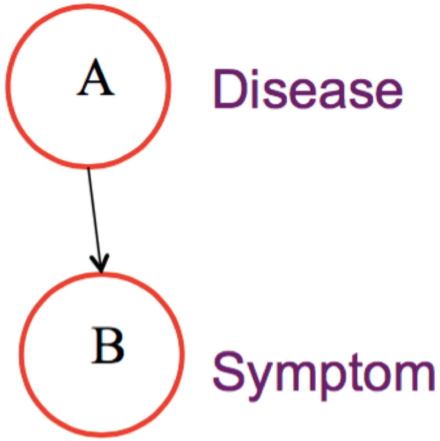
P(Symptom|Disease)

$P(B\|A)$	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Example of MAP Queries

P(Diseases)

a^0	a^1
0,4	0.6



P(Symptom|Disease)

$P(B\|A)$	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

$P(A,B) = \quad P(B|A)\,P(A)$

P(A,B)	b^0	b^1
a^0	0.04	0.36
a^1	0.3	0.3

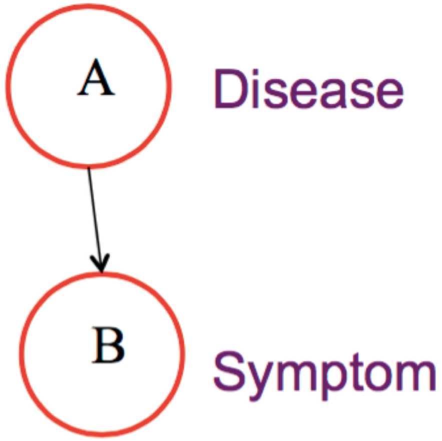
Example of MAP Queries

- Medical Diagnosis Problem
  - Diseases (*A*) cause Symptoms (*B*)
  - Two possible diseases: Mono and Flu
  - Two possible symptoms: Headache and Fever
- Q2: Most likely disease and symptom *P(A,B)*?

$$
\begin{aligned}
MAP(A,B) &= \arg\max_{a,b} P(A,B) \\
&= \arg\max_{a,b} P(B \mid A) P(A) \\
&= \arg\max_{a,b} \{0.04, 0.36, 0.3, 0.3\} \\
&= a^0, b^1
\end{aligned}
$$

P(Diseases)

a^0	a^1
0,4	0.6



P(Symptom|Disease)

P(B|A)	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Marginal MAP Query

- We looked for highest joint probability assignment of disease and symptom
- Can look for most likely assignment of disease variable only
- Query is not all remaining variables but a subset of them
  - *Y* is query, evidence is *E=e*
    Task is to find most likely assignment to *Y*:

$$MAP(Y \mid e) = \arg\max_{y} P(y \mid e)$$

  - If *Z=X-Y-E*

$$MAP(Y \mid e) = \arg\max_{y} \sum_{z} P(y, z \mid e)$$

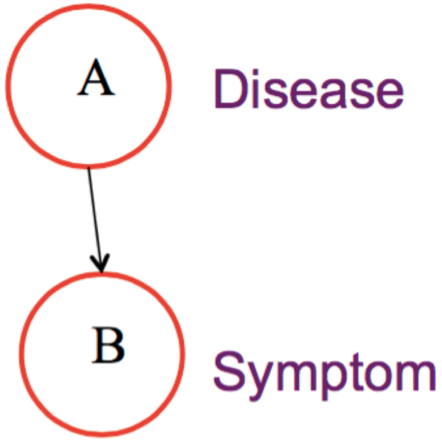
Example of MAP Queries

- Medical Diagnosis Problem
  - Diseases (A) cause Symptoms (B)
  - Two possible diseases: Mono and Flu
  - Two possible symptoms: Headache and Fever

- Q3: Most likely symptom P(B)?

P(Diseases)

a^0	a^1
0,4	0.6



P(Symptom|Disease)

$P(B\|A)$	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Example of MAP Queries

- Q3: Most likely symptom P(*B*)?

$$MAP(B) = \arg\max_b P(b) = \arg\max_b \sum_a P(a,b)$$

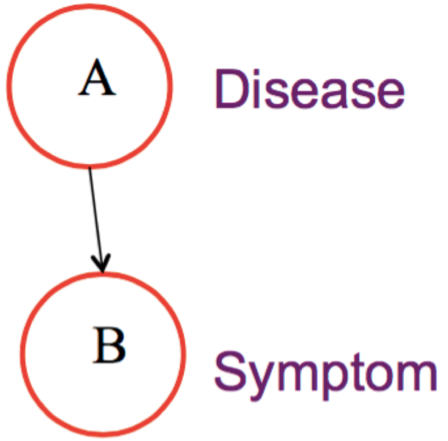
$$= \arg\max_b \{0.34, 0.66\} = b^1$$

P(A,B) = P(A)P(B|A)

P(A,B)	b^0	b^1
a^0	0.04	0.36
a^1	0.3	0.3

P(Diseases)

a^0	a^1
0,4	0.6



P(Symptom|Disease)

$P(B\|A)$	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

Marginal MAP Assignments

- They are not monotonic
- Most likely assignment $MAP(Y_1|e)$ might be completely different from assignment to Y_1 in $MAP(\{Y_1,Y_2\}|e)$
  - Q1: Most likely disease P(A)?
  - A1: Flu
  - Q2: Most likely disease and symptom P(A,B)?
  - A2: Mono and Fever
- Thus we cannot use a MAP query to give a correct answer to a marginal map query

Marginal MAP more Complex than MAP

- Contains both summations (like in probability queries) and maximizations (like in MAP queries)

$$MAP(B) = \arg\max_b P(b) = \arg\max_b \sum_a P(a, b)$$
$$= \arg\max_b\{0.34, 0.66\} = b^1$$

Linear Algebra For Machine Learning

Scalar

- Single number
- Represented in lower-case italic *x*
    - E.g., let $x \in \mathbb{R}$ be the slope of the line
        - Defining a real-valued scalar
    - E.g., let $n \in \mathbb{N}$ be the number of units
        - Defining a natural number scalar

Vector

- An array of numbers
- Arranged in order
- Each no. identified by an index
- Vectors are shown in lower-case bold
- If each element is in *R* then x is in R^n
- We think of vectors as points in space
  - Each element gives coordinate along an axis

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \Rightarrow \mathbf{x}^T = \begin{bmatrix} x_1, x_2, \dots, x_n \end{bmatrix}$$

Matrix

- 2-D array of numbers
- Each element identified by two indices
- Denoted by bold typeface ***A***
- Elements indicated as $A_{m,n}$
  - E.g.,

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

- *A*[i:] is *i*th row of *A*, *A*[:j] is *j*th column of ***A***
- If *A* has shape of height *m* and width *n* with real-values then $\mathbf{A} = \mathbb{R}^{m \times n}$

Tensor

- Sometimes need an array with more than two axes
- An array arranged on a regular grid with variable number of axes is referred to as a tensor
- Denote a tensor with bold typeface: A
- Element *(i,j,k)* of tensor denoted by $A_{i,j,k}$

Transpose of a Matrix

- Mirror image across principal diagonal

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} \Rightarrow A^T = \begin{bmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \\ x_{13} & x_{23} & x_{33} \end{bmatrix}$$

- Vectors are matrices with a single column
  - Often written in-line using transpose

$$\mathbf{x} = [x_1,..,x_n]^\mathrm{T}$$

- Since a scalar is a matrix with one element $a = a^\mathrm{T}$

Linear Transformation

$$A\mathbf{x} = \mathbf{b}$$

- where $A \in \mathrm{R}^{n \times n}$ and $\mathbf{b} \in \mathrm{R}^n$

$$A_{11}x_1 + A_{12}x_2 + ... + A_{1n}x_n = b_1$$
$$A_{21}x_1 + A_{22}x_2 + ... + A_{2n}x_n = b_2$$
$$...$$
$$A_{n1}x_1 + A_{n2}x_2 + ... + A_{nn}x_n = b_n$$

n equations in *n* unknowns

Linear Transformation

$$A\mathbf{x} = \mathbf{b}$$

- where $A \in \mathbb{R}^{n \times n}$ and $\mathbf{b} \in \mathbb{R}^n$
- More explicitly

$$A = \begin{bmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & \vdots & \vdots \\ A_{n,1} & \cdots & A_{nn} \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

$n \times n$ $\qquad$ $n \times 1$ $\qquad$ $n \times 1$

Can view A as a *linear transformation* of vector ***x*** to vector ***b***

- Sometimes we wish to solve for the unknowns $x = \{x_1, .., x_n\}$ when A and ***b*** provide constraints

Identity and Inverse Matrices

- Matrix inversion is a powerful tool to analytically solve $A\boldsymbol{x}=\boldsymbol{b}$
- Needs concept of Identity matrix
- Identity matrix does not change value of vector
- when we multiply the vector by identity matrix
  - Denote identity matrix that preserves n-dimensional vectors as I_n
  - Formally $I_n \in \mathbb{R}^{n\times n}$ and $\forall \mathbf{x} \in \mathbb{R}^n$, $I_n \mathbf{x} = \mathbf{x}$
  - Example of I_3

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Matrix Inverse

- Inverse of square matrix A defined as $A^{-1}A=I_n$
- We can now solve $A\boldsymbol{x}=\boldsymbol{b}$ as follows:

$$A\boldsymbol{x}=\boldsymbol{b}$$
$$A^{-1}A\boldsymbol{x} = A^{-1}\boldsymbol{b}$$
$$I_n\,\boldsymbol{x} = A^{-1}\boldsymbol{b}$$
$$\boldsymbol{x} = A^{-1}\boldsymbol{b}$$

- This depends on being able to find A^{-1}
- If A^{-1} exists there are several methods for finding it

Solving Simultaneous equations

- $A\boldsymbol{x} = \boldsymbol{b}$
- Two closed-form solutions
  - Matrix inversion $\boldsymbol{x}=A^{-1}\boldsymbol{b}$
  - Gaussian elimination

Norms

- Used for measuring the size of a vector
- Norms map vectors to non-negative values
- Norm of vector ***x*** is distance from origin to ***x***
  - It is any function *f* that satisfies:

$$f(\boldsymbol{x}) = 0 \Rightarrow \boldsymbol{x} = 0$$

$$f(\boldsymbol{x}+\boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y}) \quad \text{Triangle Inequality}$$

$$\forall \alpha \in \mathbb{R} \quad f(\alpha \boldsymbol{x}) = |\alpha| f(\boldsymbol{x})$$

L^P Norm

- Definition

$$||x||_p = (\sum_i |x_i|^p)^{\frac{1}{p}}$$

L^P Norm

- Definition $||x||_p = (\sum_i |x_i|^p)^{\frac{1}{p}}$
- L^2 Norm
  - Called Euclidean norm, written simply as $||x||$
  - Squared Euclidean norm is same as $x^T x$

$$||x||_2 = \sqrt{\sum_i |x_i|^2}$$

$$= \sqrt{x^T x}$$

L^P Norm

- Definition $||x||_p = (\sum_i |x_i|^p)^{\frac{1}{p}}$
- L^1 Norm
  - also called Manhattan distance

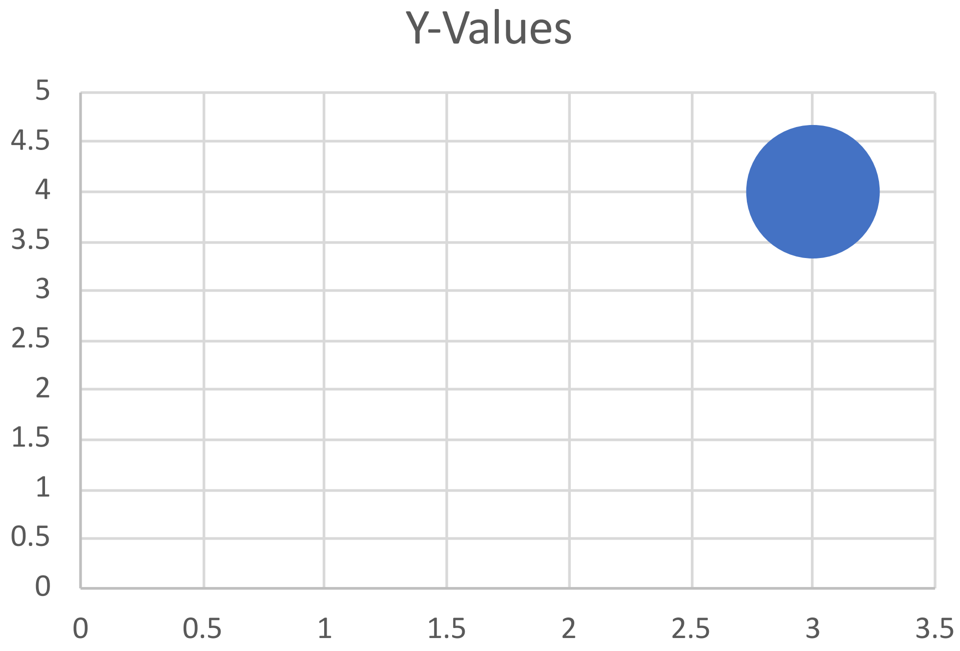
$$||x||_1 \quad = \quad \sum_i |x_i|$$

L^P Norm

- Definition $||x||_p = (\sum_i |x_i|^p)^{\frac{1}{p}}$
- L^∞ Norm
  - also called max norm

$$||x||_\infty \quad = \quad \max_i |x_i|$$

Norms of two-dimensional Point



X = (3,4)

$||x||_1$ = 3+4 = 5 $\qquad ||x||_1 = \sum_i |x_i|$

$||x||_2 = \sqrt{3^2 + 4^2} = 5 \qquad ||x||_2 = \sqrt{\sum_i |x_i|^2}$

$||x||_\infty = \max\{3,4\} = 4 \qquad ||x||_\infty = \max_i |x_i|$

Size of a Matrix

- Frobenius norm

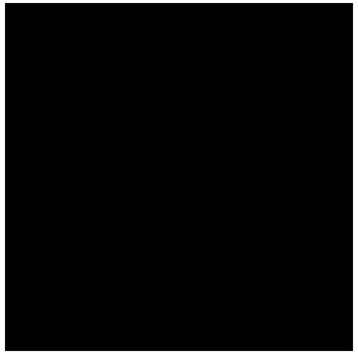
$$||A||_F = \left(\sum_{i,j} A_{i,j}^2\right)^{\frac{1}{2}}$$

- It is analogous to L^2 norm of a vector

Image distance

-

=



$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,32} \\ x_{2,1} & x_{22} & \cdots & x_{2,32} \\ \vdots & \vdots & \ddots & \vdots \\ x_{32,1} & x_{32,2} & \cdots & x_{32,32} \end{bmatrix} - \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,32} \\ y_{2,1} & y_{22} & \cdots & y_{2,32} \\ \vdots & \vdots & \ddots & \vdots \\ y_{32,1} & y_{32,2} & \cdots & y_{32,32} \end{bmatrix} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,32} \\ z_{2,1} & z_{22} & \cdots & z_{2,32} \\ \vdots & \vdots & \ddots & \vdots \\ z_{32,1} & z_{32,2} & \cdots & z_{32,32} \end{bmatrix}$$

L^1 distance between X and Y: $\sum_{i,j} |z_{i,j}| = \sum_{i,j} |x_{i,j} - y_{i,j}|$

L^2 distance between X and Y: $\sqrt{\sum_{i,j} z_{i,j}^2} = \sqrt{\sum_{i,j} (x_{i,j} - y_{i,j})^2}$

L^∞ distance between X and Y:

$$\max_{i,j} |z_{i,j}| = \max_{i,j} |x_{i,j} - y_{i,j}|$$