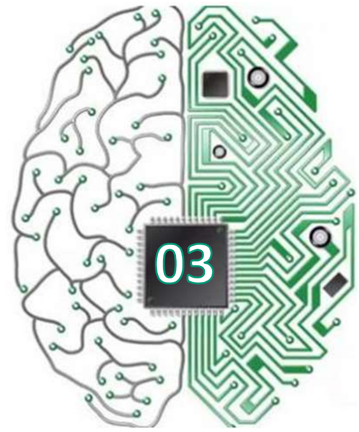
Open Elective Course [OE]
Course Code: CSO507
Winter 2023-24

Lecture#



Deep Learning

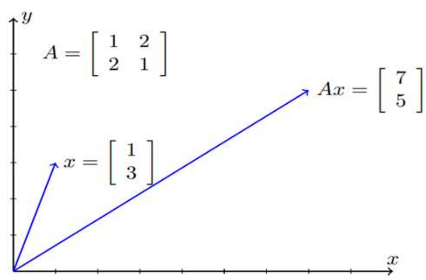
Unit-1: Linear Algebra & Vector Calculus for ML/DL
Probability Theory

Course Instructor:

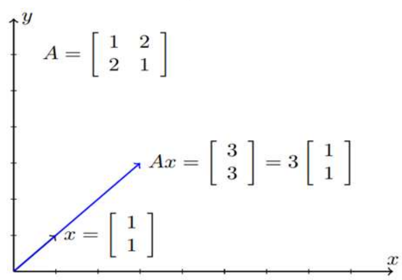
Dr. Monidipa Das
Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Technology (Indian School of Mines) Dhanbad, Jharkhand 826004, India

Eigenvectors

- When a matrix hits a vector, the vector gets transformed into a new vector (it strays from its path). The vector may also get scaled (elongated or shortened) in the process.



- **For a given square matrix A, there exist special vectors which refuse to stray from their path.** These vectors are called **eigenvectors.**



Eigen Decomposition

3

- ***Eigen decomposition*** is decomposing a matrix into a set of eigenvalues and eigenvectors
- ***Eigenvalues*** of a square matrix $\mathbf{A}$ are scalars λ and ***eigenvectors*** are non-zero vectors $\mathbf{v}$ that satisfy
$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$
- Eigenvalues are found by solving the following equation
$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$
- **Eigenvalues and eigenvectors are only for square matrices.**
- Eigenvectors are *by definition nonzero*. Eigenvalues may be equal to zero.
- **We do not consider the zero vector to be an eigenvector:** since A0=0=λ0 for *every* scalar λ, the associated eigenvalue would be undefined.
- Several properties of matrices can be analyzed based on their eigenvalues
- The eigenvectors of a matrix $A \in \mathbb{R}^{n \times n}$ having distinct eigenvalues are linearly independent.
- The eigenvectors of a square symmetric matrix are orthogonal.

Eigen Decomposition

4

- If a matrix $\mathbf{A}$ has n linearly independent eigenvectors $\{\mathbf{v}^1, \dots, \mathbf{v}^n\}$ with corresponding eigenvalues $\{\lambda_1, \dots, \lambda_n\}$, the eigen decomposition of $\mathbf{A}$ is given by
$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$$
  - Columns of the matrix $\mathbf{V}$ are the eigenvectors, i.e., $\mathbf{V} = [\mathbf{v}^1, \dots, \mathbf{v}^n]$
  - $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues, i.e.,$\mathbf{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_n]$
- To find the inverse of the matrix A, we can use $\mathbf{A}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^{-1}$
  - This involves simply finding the inverse $\mathbf{\Lambda}^{-1}$ of a diagonal matrix

Eigen Decomposition

5

- Decomposing a matrix into eigenvalues and eigenvectors allows to analyze certain properties of the matrix
  - If all eigenvalues are positive, the matrix is positive definite
  - If all eigenvalues are positive or zero-valued, the matrix is positive semidefinite
  - If all eigenvalues are negative or zero-values, the matrix is negative semidefinite
    - Positive semidefinite matrices are interesting because they guarantee that $\forall \mathbf{x}, \mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$
- Eigen decomposition can also simplify many linear-algebraic computations
  - The determinant of A can be calculated as

$$\det(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$$

  - **If any of the eigenvalues are zero, the matrix is singular (it does not have an inverse)**
- However, not every matrix can be decomposed into eigenvalues and eigenvectors
  - Also, in some cases the decomposition may involve complex numbers
  - Still, every real symmetric matrix is guaranteed to have an eigen decomposition according to $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, where $\mathbf{V}$ is an orthogonal matrix

Singular Value Decomposition

6

- ***Singular value decomposition*** (SVD) provides another way to factorize a matrix, into singular vectors and singular values
  - SVD is more generally applicable than eigen decomposition
  - Every real matrix has an SVD, but the same is not true of the eigen decomposition
    - E.g., if a matrix is not square, the eigen decomposition is not defined, and we must use SVD
- SVD of an $m \times n$ matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^{\boldsymbol{T}}$$

  - $\mathbf{U}$ is an $m \times m$ matrix, $\mathbf{D}$ is an $m \times n$ matrix, and $\mathbf{V}$ is an $n \times n$ matrix
  - The elements along the diagonal of $\mathbf{D}$ are known as the singular values of A
  - The columns of $\mathbf{U}$ are known as the left-singular vectors
  - The columns of $\mathbf{V}$ are known as the right-singular vectors
- For a non-square matrix $\mathbf{A}$, the squares of the singular values σ_i are the eigenvalues λ_i of $\mathbf{A}^{\boldsymbol{T}}\mathbf{A}$, i.e., $\sigma_i^2 = \lambda_i$ for $i = 1,2, \dots, n$

Prof. Monidipa Das, Department of CSE, IIT (ISM) Dhanbad

Vector Calculus

Differential Calculus

- For a function $f: \mathbb{R} \to \mathbb{R}$, the ***derivative*** of f is defined as

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

- If $f'(a)$ exists, f is said to be differentiable at a
- If $f\,‘(c)$ exits for $\forall c \in [a, b]$, then f is differentiable on this interval
  - We can also interpret the derivative $f'(x)$ as the instantaneous rate of change of $f(x)$ with respect to *x*
  - I.e., for a small change in *x*, what is the rate of change of $f(x)$
- Given $y = f(x)$, where *x* is an independent variable and *y* is a dependent variable, the following expressions are equivalent:

$$f'(x) = f' = \frac{dy}{dx} = \frac{df}{dx} = \frac{d}{dx} f(x) = Df(x) = D_x f(x)$$

- The symbols $\frac{d}{dx}$, *D*, and D_x are differentiation operators that indicate operation of differentiation

Differential Calculus

9

- The following rules are used for computing the derivatives of explicit functions
  - **Derivative of constants.** $\frac{d}{dx}c = 0$.
  - **Derivative of linear functions.** $\frac{d}{dx}(ax) = a$.
  - **Power rule.** $\frac{d}{dx}x^n = nx^{n-1}$.
  - **Derivative of exponentials.** $\frac{d}{dx}e^x = e^x$.
  - **Derivative of the logarithm.** $\frac{d}{dx}\log(x) = \frac{1}{x}$.
  - **Sum rule.** $\frac{d}{dx}(g(x) + h(x)) = \frac{dg(x)}{dx} + \frac{dh(x)}{dx}$.
  - **Product rule.** $\frac{d}{dx}(g(x) \cdot h(x)) = g(x)\frac{dh(x)}{dx} + \frac{dg(x)}{dx}h(x)$.
  - **Chain rule.** $\frac{d}{dx}g(h(x)) = \frac{dg(h(x))}{dh} \cdot \frac{dh(x)}{dx}$.

Matrix Calculus

10

- **Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ then** $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} = \boldsymbol{A}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$
- Let $\alpha = \boldsymbol{y}^T\boldsymbol{x}$ then $\frac{\partial \alpha}{\partial \boldsymbol{z}} = \boldsymbol{x}^T\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} + \boldsymbol{y}^T\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$
- Let $\alpha = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x}$ **then** $\frac{\partial \alpha}{\partial \boldsymbol{z}} = \boldsymbol{x}^T\boldsymbol{A}^T\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} + \boldsymbol{y}^T\boldsymbol{A}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$
- Let $\alpha = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ **then** $\frac{\partial \alpha}{\partial \boldsymbol{z}} = \boldsymbol{x}^T(\boldsymbol{A}^T + \boldsymbol{A})\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$
- **Let A be symmetric and $\alpha = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ then** $\frac{\partial \alpha}{\partial \boldsymbol{z}} = 2\boldsymbol{x}^T\boldsymbol{A}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$

Higher Order Derivatives

11

- The derivative of the first derivative of a function $f(x)$ is the ***second derivative*** of $f(x)$

$$\frac{d^2 f}{dx^2} = \frac{d}{dx}\left(\frac{df}{dx}\right)$$

- The second derivative quantifies how the rate of change of $f(x)$ is changing
  - E.g., in physics, if the function describes the displacement of an object, the first derivative gives the velocity of the object (i.e., the rate of change of the position)
  - The second derivative gives the acceleration of the object (i.e., the rate of change of the velocity)

- If we apply the differentiation operation any number of times, we obtain the *n*-th derivative of $f(x)$

$$f^{(n)}(x) = \frac{d^n f}{dx^n} = \left(\frac{d}{dx}\right)^n f(x)$$

Partial Derivatives

12

- Functions that depend on many variables are called multivariate functions
- Let $y = f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$ be a multivariate function with *n* variables
  - The input is an *n*-dimensional vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ and the output is a scalar *y*
  - The mapping is $f: \mathbb{R}^n \rightarrow \mathbb{R}$

- The ***partial derivative*** of *y* with respect to its *i*th parameter x_i is

$$\frac{\partial y}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, x_2, \ldots, x_i + h, \ldots, x_n) - f(x_1, x_2, \ldots, x_i, \ldots, x_n)}{h}$$

- To calculate $\frac{\partial y}{\partial x_i}$ (∂ pronounced "del" or we can just say "partial derivative"), we can treat $x_1, x_2, \ldots, x_{i-1}, \; x_{i+1} \ldots, x_n$ as constants and calculate the derivative of *y* only with respect to x_i

- For notation of partial derivatives, the following are equivalent:

$$\frac{\partial y}{\partial x_i} = \frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i} f(\mathbf{x}) = f_{x_i} = f_i = D_i f = D_{x_i} f$$

Gradient

13

- We can concatenate partial derivatives of a multivariate function with respect to all its input variables to obtain the ***gradient*** vector of the function
- The gradient of the multivariate function $f(\mathbf{x})$ with respect to the *n*-dimensional input vector $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$, is a vector of *n* partial derivatives

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right]^T$$

- When there is no ambiguity, the notations $\nabla f(\mathbf{x})$ or $\nabla_{\mathbf{x}} f$ are often used for the gradient instead of $\nabla_{\mathbf{x}} f(\mathbf{x})$
  - The symbol for the gradient is the Greek letter ∇ (pronounced "nabla"), although $\nabla_{\mathbf{x}} f(\mathbf{x})$ is more often it is pronounced "gradient of *f* with respect to **x**"
- In ML, the gradient descent algorithm relies on the opposite direction of the gradient of the loss function $\mathcal{L}$ with respect to the model parameters θ ($\nabla_\theta \mathcal{L}$) for minimizing the loss function
  - Adversarial examples can be created by adding perturbation in the direction of the gradient of the loss $\mathcal{L}$ with respect to input examples x ($\nabla_x \mathcal{L}$) for maximizing the loss function

Hessian Matrix

14

- To calculate the second-order partial derivatives of multivariate functions, we need to calculate the derivatives for all combination of input variables
- That is, for a function $f(\mathbf{x})$ with an *n*-dimensional input vector $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$, there are n^2 second partial derivatives for any choice of *i* and *j*

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)$$

- The second partial derivatives are assembled in a matrix called the ***Hessian***

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}$$

- Computing and storing the Hessian matrix for functions with high-dimensional inputs can be computationally prohibitive
  - E.g., the loss function for a ResNet50 model with approximately 23 million parameters, has a Hessian of 23 M × 23 M = 529 T (trillion) parameters

Jacobian Matrix

- The concept of derivatives can be further generalized to vector-valued functions (or, vector fields) $f: \mathbb{R}^n \to \mathbb{R}^m$
- For an n-dimensional input vector $\mathbf{x} = [x_1, x_2, \dots, x_n]^T \in \mathbb{R}^n$, the vector of functions is given as

$$\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x})]^T \in \mathbb{R}^m$$

- The matrix of first-order partial derivates of the vector-valued function $\mathbf{f}(\mathbf{x})$ is an $m \times n$ matrix called a ***Jacobian***

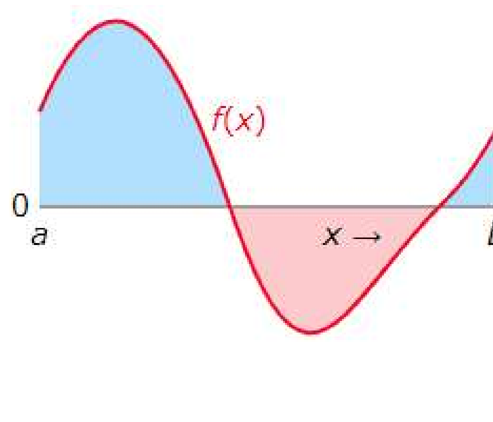
$$\mathbf{J} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

  - For example, in robotics a robot Jacobian matrix gives the partial derivatives of the translational and angular velocities of the robot end-effector with respect to the joints (i.e., axes) velocities

Integral Calculus

- For a function $f(x)$ defined on the domain $[a, b]$, the definite ***integral*** of the function is denoted

$$\int_a^b f(x)dx$$



- Geometric interpretation of the integral is the area between the horizontal axis and the graph of $f(x)$ between the points *a* and *b*
  - In this figure, the integral is the sum of blue areas (where $f(x) > 0$) minus the pink area (where $f(x) < 0$)

Probability Theory

Probability

- Intuition:
  - In a process, several outcomes are possible
  - When the process is repeated a large number of times, each outcome occurs with a ***relative frequency***, or ***probability***
  - If a particular outcome occurs more often, we say it is more probable
- Probability arises in two contexts
  - In actual repeated experiments
    - Example: You record the color of 1,000 cars driving by. 57 of them are green. You estimate the probability of a car being green as 57/1,000 = 0.057.
  - In idealized conceptions of a repeated process
    - Example: You consider the behavior of an unbiased six-sided die. The expected probability of rolling a 5 is 1/6 = 0.1667.
    - Example: You need a model for how people's heights are distributed. You choose a normal distribution to represent the expected relative probabilities.

Probability

19

- Solving machine learning problems requires to deal with uncertain quantities, as well as with stochastic (non-deterministic) quantities
  - Probability theory provides a mathematical framework for representing and quantifying uncertain quantities
- There are different sources of uncertainty:
  - **Inherent stochasticity in the system being modeled**
    - For example, most interpretations of quantum mechanics describe the dynamics of subatomic particles as being probabilistic
  - **Incomplete observability**
    - Even deterministic systems can appear stochastic when we cannot observe all of the variables that drive the behavior of the system
  - **Incomplete modeling**
    - When we use a model that must discard some of the information we have observed, the discarded information results in uncertainty in the model's predictions
    - E.g., discretization of real-numbered values, dimensionality reduction, etc.

Axioms of probability

20

- The probability of an event $\mathcal{A}$ in the given sample space $\mathcal{S}$, denoted as $P(\mathcal{A})$, must satisfies the following properties:
  - Non-negativity
    - For any event $\mathcal{A} \in \mathcal{S}$, $P(\mathcal{A}) \geq 0$
  - All possible outcomes
    - Probability of the entire sample space is 1, $P(\mathcal{S}) = 1$
  - Additivity of disjoint events
    - For all events $\mathcal{A}_1, \mathcal{A}_2 \in \mathcal{S}$ that are mutually exclusive ($\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$), the probability that both events happen is equal to the sum of their individual probabilities, $P(\mathcal{A}_1 \cup \mathcal{A}_2) = P(\mathcal{A}_1)+P(\mathcal{A}_2)$

Random variables

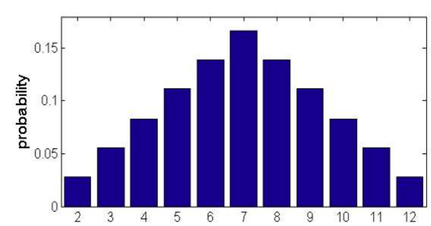
21

- A ***random variable*** X is a variable that can take on different values
  - Example: X = rolling a die
    - Possible values of X comprise the **sample space**, or **outcome space**, $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
    - We denote the event of "seeing a 5" as $\{X = 5\}$ or $X = 5$
    - The probability of the event is $P(\{X = 5\})$ or $P(X = 5)$
    - Also, $P(5)$ can be used to denote the probability that X takes the value of 5
- The probability of a random variable $P(X)$ must obey the axioms of probability over the possible values in the sample space $\mathcal{S}$
- A ***probability distribution*** is a description of how likely a random variable is to take on each of its possible states
  - A compact notation is common, where $P(X)$ is the probability distribution over the random variable X
    - Also, the notation $X \sim P(X)$ can be used to denote that the random variable X has probability distribution $P(X)$
- Random variables can be discrete or continuous
  - Discrete random variables have finite number of states: e.g., the sides of a die
  - Continuous random variables have infinite number of states: e.g., the height of a person

Discrete Variables

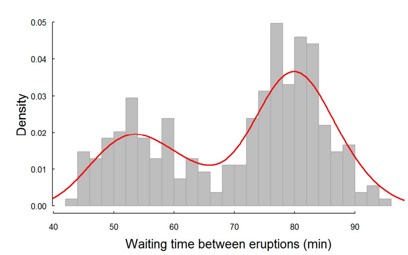
22

- A probability distribution over discrete variables may be described using a ***probability mass function*** (PMF)
  - E.g., sum of two dice



- A probability distribution over continuous variables may be described using a ***probability density function*** (PDF)
  - A PDF gives the probability of an infinitesimal region with volume δX
  - To find the probability over an interval $[a, b]$, we can integrate the PDF as follows:

$$P(X \in [a, b]) = \int_a^b P(X)dX$$



Multivariate Random Variables

23

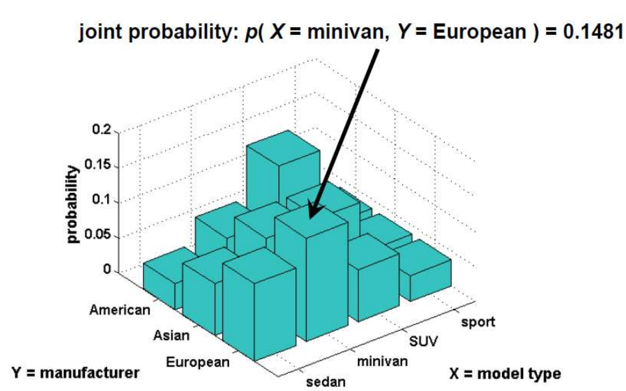
- We may need to consider several random variables at a time
  - If several random processes occur in parallel or in sequence
  - E.g., to model the relationship between several diseases and symptoms
  - E.g., to process images with millions of pixels (each pixel is one random variable)
- Probability distributions defined over multiple random variables
  - These include joint and conditional probability distributions
- The individual random variables can also be grouped together into a random vector, because they represent different properties of an individual statistical unit
- A ***multivariate random variable*** is a vector of multiple random variables

$$\mathbf{X} = [X_1, X_2, \dots, X_n]^T$$

Joint Probability Distribution

24

- Probability distribution that acts on many variables at the same time is known as a ***joint probability distribution***
- Given any values *x* and *y* of two random variables X and Y, what is the probability that X = *x* and Y = *y* simultaneously?
  - $P(X = x, Y = y)$ denotes the joint probability
  - We may also write $P(x, y)$ for brevity



Marginal Probability Distribution

25

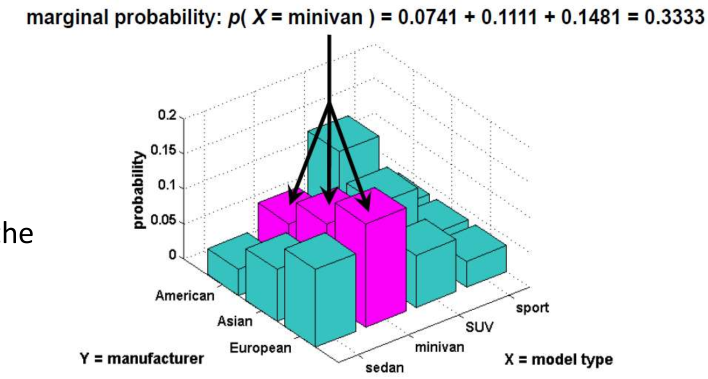
- ***Marginal probability distribution*** is the probability distribution of a single variable
  - It is calculated based on the joint probability distribution $P(X, Y)$
  - I.e., using the sum rule:

$$P(X=x)=\sum_{y} P(X=x, Y=y)$$

  - For continuous random variables, the summation is replaced with integration,

$$P(X=x)=\int P(X=x, Y=y)\, dy$$

  - This process is called marginalization



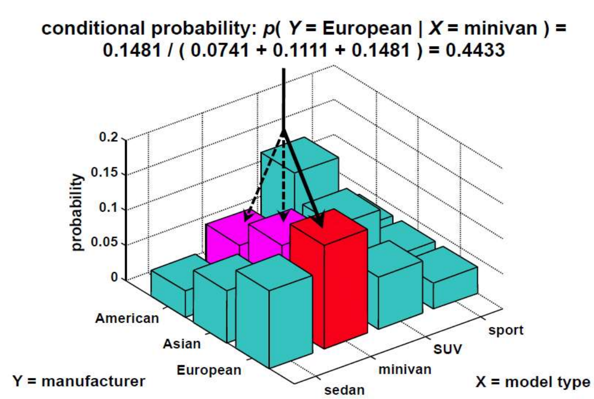
Conditional Probability Distribution

26

- ***Conditional probability distribution*** is the probability distribution of one variable provided that another variable has taken a certain value
  - Denoted $P(X=x|\, Y=y)$

- Note that:

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



Exercise

27

1. **Suppose $f(x) = sin^3(x)$. Then, $f'(x) = ?$**
   A. $f'(x) = \cos^3(x)$
   B. $f'(x) = 3\cos^2(x)$
   C. $f'(x) = 3\sin^2(x)$
   D. $f'(x) = 3\sin^2(x)\cos(x)$

2. **Suppose $f(x) = (x+2)(3x-3)$. Then, $f'(x) = ?$**
   A. $f'^{(x)} = 3$
   B. $f'^{(x)} = 3x + 6$
   C. $f'^{(x)} = 6x + 3$
   D. $None\ of\ these$

3. **If $f(x, y) = x^3y + x + 2y$. Then the Hessian matrix would be of dimension:**
   A. 2×2
   B. 2×1
   C. 4×4
   D. 1×2

4. **If $f(x, y) = x^3y + x + 2y$. Then the Hessian matrix would be:**

   ? ____________________

5. **Suppose $\mathbf{f}(x, y) = [x^2 - y^2, 2xy]^T$. Then $J(x, y) = ?$** [$J(x, y)$ indicates Jacobian of $f(x, y)$]

   ? ________________________

28

Questions?

Acknowledgement:
- Dr. Alex Vakanski