School of Computing
UNIVERSITY OF GEORGIA

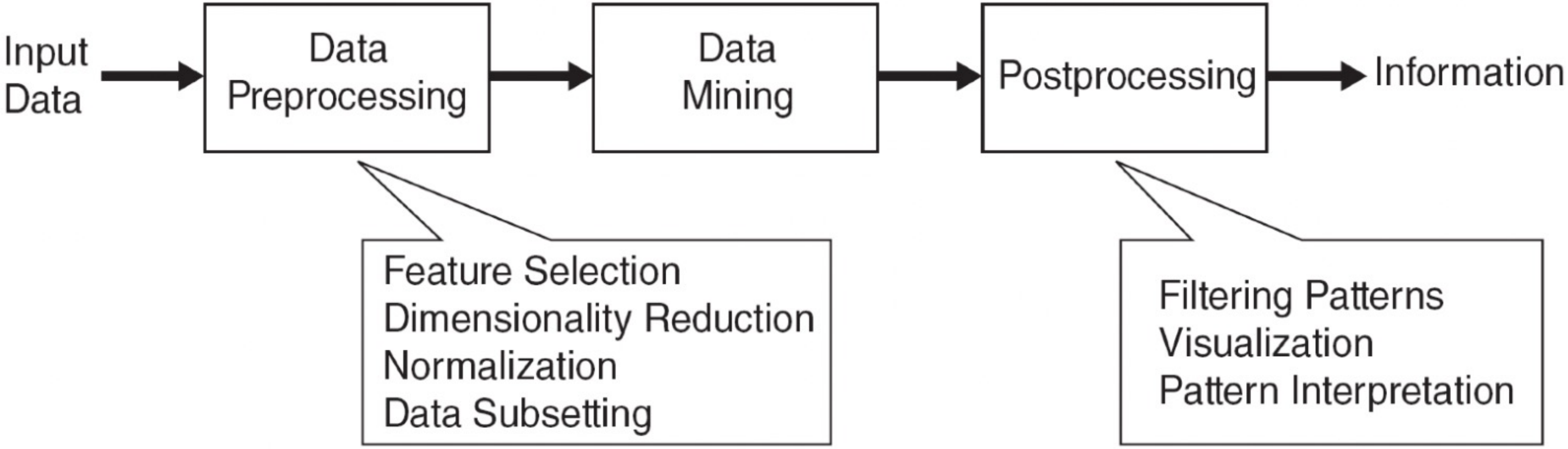
CSCI 4380/6380 Data Mining

Fei Dou

Assistant Professor
School of Computing
University of Georgia

August 29, 2023

Recap: Data Mining Process



Recap

- Google Colab
- Python Basics
- NumPy
- Pandas
- Matplotlib
- *PyCharm/Visual Studio + Anaconda

Course Details

- **TA:** Yucheng Shi (yucheng.shi@uga.edu)
- **Office Hours:**
  - Mon: 12:30 PM - 1:30 PM; Tue: 11:30 AM – 12: 30 PM; Thu: 11:30 AM – 12: 30 PM
    - Mon: https://zoom.us/j/93255326212?pwd=cXZQOGo5TUlIMC9Rb0w1Ny9jbytGUT09
    - Tue&Thu: https://zoom.us/j/98101155911?pwd=MWJDUGh5NVUzVW1oa0NVaFhYOWs2QT09

  - Also, by appointment at Boyd 307

  - Location:
    - Boyd 307 Boyd Research and Education Center
    - Zoom Link

Mathematics for Data Mining

Lecture Outline

- Linear algebra
  - Vectors
  - Matrices
  - Eigen decomposition
- Differential calculus
- Probability
  - Random variables
  - Probability theory

Linear algebra

Notation

- a, b, c Scalar (integer or real)
- $\mathbf{x}, \mathbf{y}, \mathbf{z}$ Vector (bold-font, lower case)
- $\mathbf{A}, \mathbf{B}, \mathbf{C}$ Matrix (bold-font, upper-case)
- $\mathsf{\mathbf{A}}, \mathsf{\mathbf{B}}, \mathsf{\mathbf{C}}$ Tensor ((bold-font, upper-case)
- X, Y, Z Random variable (normal font, upper-case)
- $a \in \mathcal{A}$ Set membership: a is member of set $\mathcal{A}$
- $|\mathcal{A}|$ Cardinality: number of items in set $\mathcal{A}$
- $||\mathbf{v}||$ Norm of vector $\mathbf{v}$
- $\mathbf{u} \cdot \mathbf{v}$ or $\langle \mathbf{u}, \mathbf{v} \rangle$ Dot product of vectors $\mathbf{u}$ and $\mathbf{v}$
- $\mathbb{R}$ Set of real numbers
- $\mathbb{R}^n$ Real numbers space of dimension n
- $\mathbb{R}^{m \times n}$ Real numbers matrices of dimension m by n
- $y = f(x)$ or $x \mapsto f(x)$ Function (map): assign a unique value $f(x)$ to each input value x
- $f: \mathbb{R}^n \to \mathbb{R}$ Function (map): map an n-dimensional vector into a scalar

Notation

- $\mathbf{A} \odot \mathbf{B}$ — Element-wise product of matrices $\mathbf{A}$ and $\mathbf{B}$
- $\mathbf{A}^{\dagger}$ — Pseudo-inverse of matrix $\mathbf{A}$
- $\frac{d^n f}{dx^n}$ — n-th derivative of function f with respect to x
- $\nabla_{\mathbf{x}} f(\mathbf{x})$ — Gradient of function f with respect to $\mathbf{x}$
- $\mathbf{H}_f$ — Hessian matrix of function f
- $X \sim P$ — Random variable X has distribution P
- $P(X|Y)$ — Probability of X given Y
- $\mathcal{N}(\mu, \sigma^2)$ — Gaussian distribution with mean μ and variance σ^2
- $\mathbb{E}_{X\sim P}[f(X)]$ — Expectation of $f(X)$ with respect to $P(X)$
- $\mathrm{Var}\big(f(X)\big)$ — Variance of $f(X)$
- $\mathrm{Cov}\big(f(X), g(Y)\big)$ — Covariance of $f(X)$ and $g(Y)$
- $\mathrm{corr}(X, Y)$ — Correlation coefficient for X and Y
- $D_{KL}(P||Q)$ — Kullback-Leibler divergence for distributions P and Q
- $CE(P, Q)$ — Cross-entropy for distributions P and Q

Scalars, Vectors

- ***Scalars***: $s \in \mathbb{R}$ and $n \in \mathbb{N}$

- ***Vector*** definition
  - **Computer science**: *vector* is a one-dimensional array of ordered real-valued scalars
  - **Mathematics**: *vector* is a quantity possessing both magnitude and direction, represented by an arrow indicating the direction, and the length of which is proportional to the magnitude
- Vectors are written in column form or in row form
  - Denoted by bold-font lower-case letters

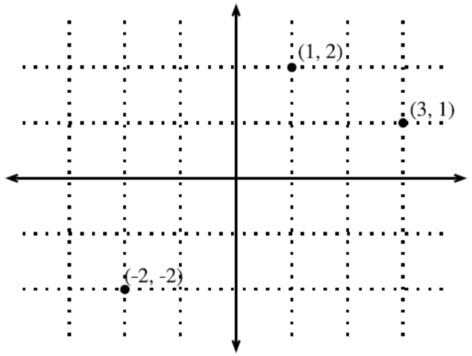
$$\mathbf{x} = \begin{bmatrix} 1 \\ 7 \\ 0 \\ 1 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} 1 & 7 & 0 & 1 \end{bmatrix}^T$$

- For a general form vector with n elements, the vector lies in the n-dimensional space $\mathbf{x} \in \mathbb{R}^n$

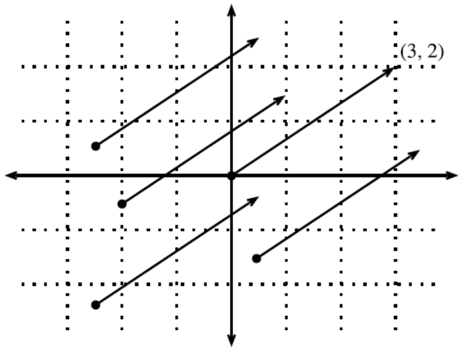
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Geometry of Vectors

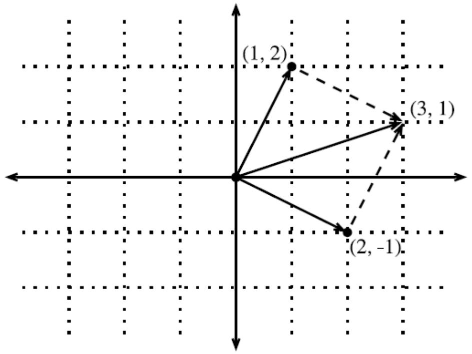
- First interpretation of a vector: point in space
  - E.g., in 2D we can visualize the data points with respect to a coordinate origin



- Second interpretation of a vector: direction in space
  - E.g., the vector $\vec{\mathbf{v}} = [3, 2]^T$ has a direction of 3 steps to the right and 2 steps up
  - The notation $\vec{\mathbf{v}}$ is sometimes used to indicate that the vectors have a direction
  - All vectors in the figure have the same direction

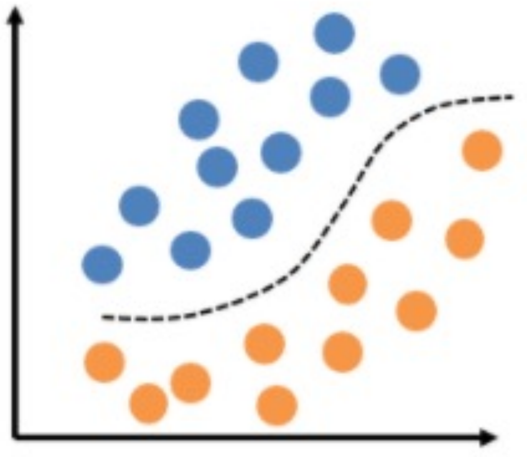


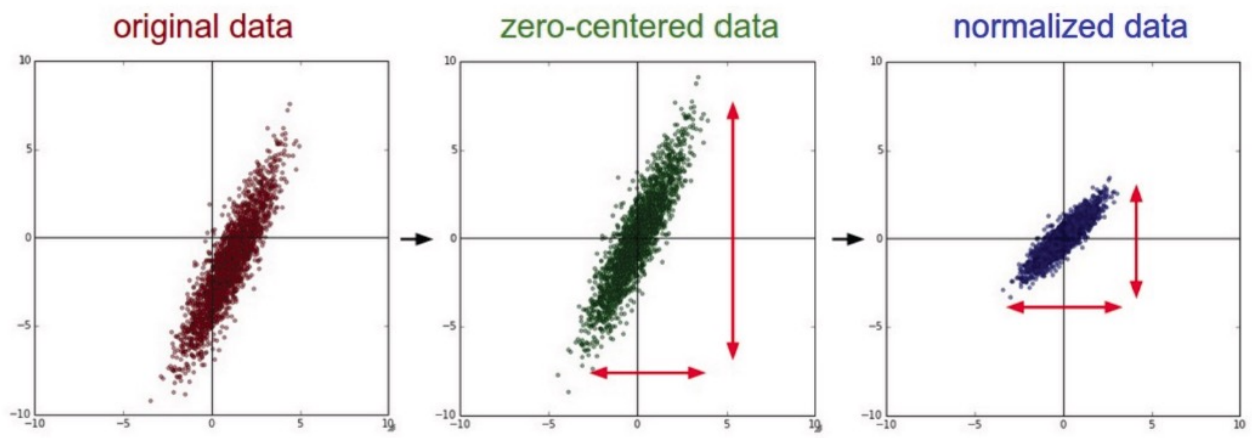
- Vector addition
  - We add the coordinates, and follow the directions given by the two vectors that are added



Geometry of Vectors

- The geometric interpretation of vectors as points in space allow us to consider a training set of input examples in ML as a collection of points in space
  - Hence, classification can be viewed as discovering how to separate two clusters of points belonging to different classes (left picture)
    - Rather than distinguishing images containing cars, planes, buildings, for example
  - Or, it can help to visualize zero-centering and normalization of training data (right picture)



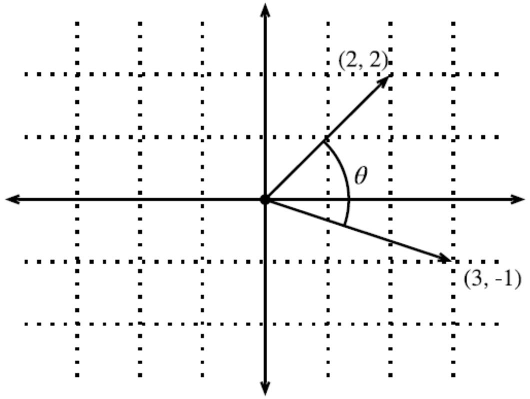


Dot Product and Angles

- ***Dot product*** of vectors, $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = \sum_i u_i \cdot v_i$
  - It is also referred to as inner product, or scalar product of vectors
  - The dot product $\mathbf{u} \cdot \mathbf{v}$ is also often denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$
- The dot product is a symmetric operation, $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = \mathbf{v}^T \mathbf{u} = \mathbf{v} \cdot \mathbf{u}$

- Geometric interpretation of a dot product:
  angle between two vectors
  - i.e., dot product $\mathbf{v} \cdot \mathbf{w}$ over the norms of the vectors is $\cos(\theta)$

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| cos(\theta) \qquad \cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$$



- If two vectors are orthogonal: $\theta = 90°$, i.e., $\cos(\theta) = 0$, then $\mathbf{u} \cdot \mathbf{v} = 0$
- Also, in ML the term $\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$ is sometimes employed as a measure of closeness of two vectors/data instances, and it is referred to as cosine similarity

Norm of a Vector

- A vector ***norm*** is a function that maps a vector to a scalar value
  - The norm is a measure of the size of the vector
- The norm f should satisfy the following properties:
  - Scaling: $f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$
  - Triangle inequality: $f(\mathbf{x}+\mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$
  - Must be non-negative: $f(\mathbf{x}) \geq 0$
- The general ℓ_p norm of a vector $\mathbf{x}$ is obtained as: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$ $\quad p \in \mathbb{R}, p > 1$
  - On next page we will review the most common norms, obtained for $p = 1, 2,$ and ∞

Norm of a Vector

- For $p = 2$, we have ℓ_2 norm
  - Also called **Euclidean norm**
  - It is the most often used norm
  - ℓ_2 norm is often denoted just as $\|\mathbf{x}\|$ with the subscript 2 omitted
  - Squared ℓ_2 norm is often used and can be obtained with $\mathbf{x}^T\mathbf{x}$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$$

- For $p = 1$, we have ℓ_1 norm
  - Uses the absolute values of the elements
  - Discriminate between zero and non-zero elements
  - L1 norm is commonly used to encourage sparsity.

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

- For $p = \infty$, we have ℓ_∞ norm
  - Known as **infinity norm**, or **max norm**
  - Outputs the absolute value of the largest element

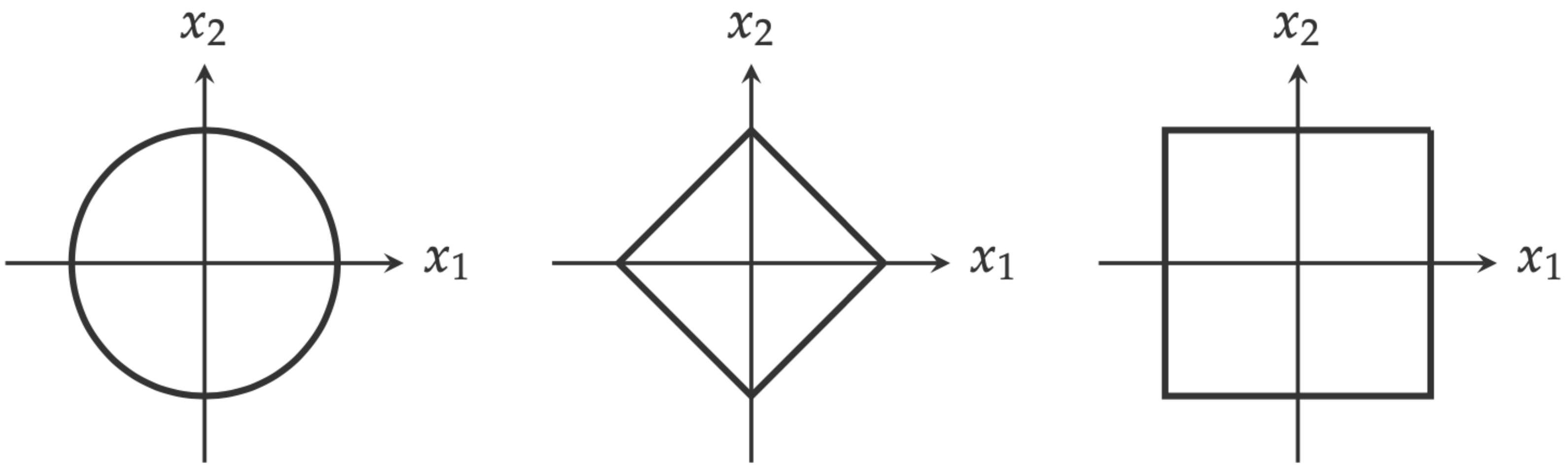
$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

- ℓ_0 norm outputs the number of non-zero elements
  - It is not an ℓ_p norm, and it is not really a norm function either (it is incorrectly called a norm)

Quiz

- For a two-dimensional vector $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$, which of the following plot is $\|\boldsymbol{x}\|_1$?

Hint: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$

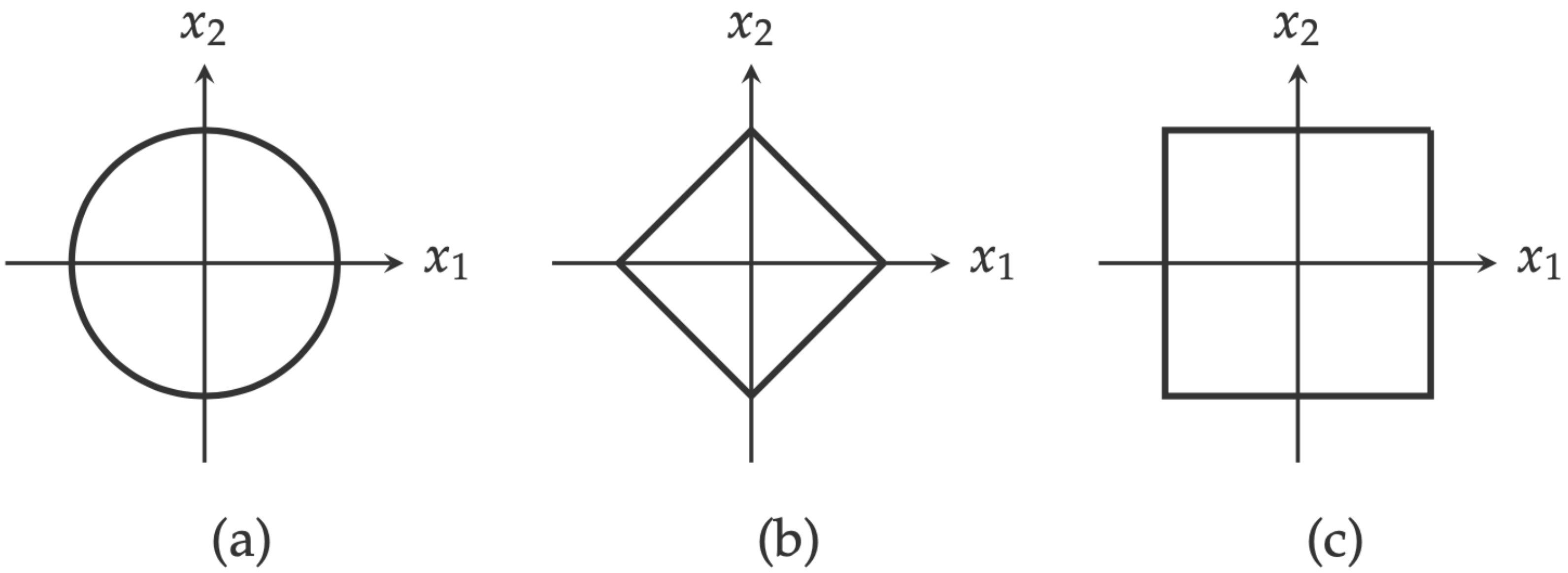


(a) (b) (c)

Quiz

- For a two-dimensional vector $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$, which of the following plot is $\|\boldsymbol{x}\|_1$?
  - Answer: (b)

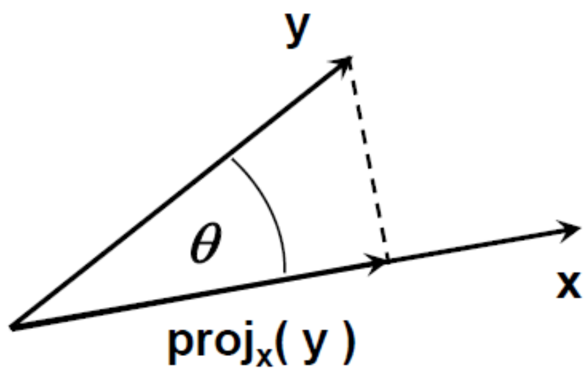
Hint: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$



Vector Projection

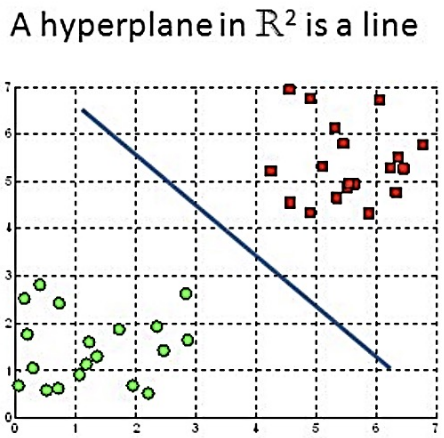
- ***Orthogonal projection*** of a vector $\mathbf{y}$ onto vector $\mathbf{x}$
  - The projection can take place in any space of dimensionality ≥ 2
  - The unit vector in the direction of $\mathbf{x}$ is $\frac{\mathbf{x}}{\|\mathbf{x}\|}$
    - A unit vector has norm equal to 1
  - The length of the projection of $\mathbf{y}$ onto $\mathbf{x}$ is $\|\mathbf{y}\| \cdot cos(\theta)$
  - The orthogonal project is the vector $\mathbf{proj}_\mathbf{x}(\mathbf{y})$

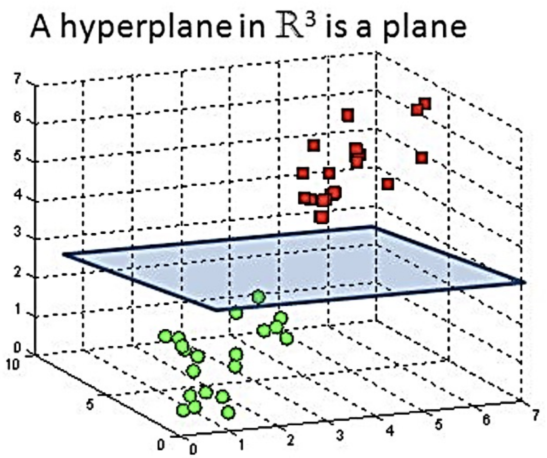
$$\mathbf{proj}_\mathbf{x}(\mathbf{y}) = \frac{\mathbf{x} \cdot \|\mathbf{y}\| \cdot cos(\theta)}{\|\mathbf{x}\|}$$



Hyperplanes

- ***Hyperplane*** is a subspace whose dimension is one less than that of its ambient space
  - In a 2D space, a hyperplane is a straight line (i.e., 1D)
  - In a 3D, a hyperplane is a plane (i.e., 2D)
  - In a d-dimensional vector space, a hyperplane has $d - 1$ dimensions, and divides the space into two half-spaces
- Hyperplane is a generalization of a concept of plane in high-dimensional space
- In ML, hyperplanes are decision boundaries used for linear classification
  - Data points falling on either sides of the hyperplane are attributed to different classes





Matrices

- ***Matrix*** is a rectangular array of real-valued scalars arranged in m horizontal rows and n vertical columns
  - Each element a_{ij} belongs to the i^{th} row and j^{th} column
  - The elements are denoted a_{ij} or $\mathbf{A}_{ij}$ or $[\mathbf{A}]_{ij}$ or $\mathbf{A}(\boldsymbol{i},\boldsymbol{j})$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}$$

- For the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the size (dimension) is $m\times n$ or (m,n)
  - Matrices are denoted by bold-font upper-case letters

Matrices

- Addition or subtraction $(\mathbf{A}\pm\mathbf{B})_{i,j} = \mathbf{A}_{i,j} \pm \mathbf{B}_{i,j}$

$$\begin{bmatrix}1 & 3 & 1\\ 1 & 0 & 0\end{bmatrix} + \begin{bmatrix}0 & 0 & 5\\ 7 & 5 & 0\end{bmatrix} = \begin{bmatrix}1+0 & 3+0 & 1+5\\ 1+7 & 0+5 & 0+0\end{bmatrix} = \begin{bmatrix}1 & 3 & 6\\ 8 & 5 & 0\end{bmatrix}$$

- Scalar multiplication $(c\mathbf{A})_{i,j} = c \cdot \mathbf{A}_{i,j}$

$$2 \cdot \begin{bmatrix}1 & 8 & -3\\ 4 & -2 & 5\end{bmatrix} = \begin{bmatrix}2\cdot 1 & 2\cdot 8 & 2\cdot -3\\ 2\cdot 4 & 2\cdot -2 & 2\cdot 5\end{bmatrix} = \begin{bmatrix}2 & 16 & -6\\ 8 & -4 & 10\end{bmatrix}$$

- Matrix multiplication $(\mathbf{AB})_{i,j} = \mathbf{A}_{i,1}\mathbf{B}_{1,j} + \mathbf{A}_{i,2}\mathbf{B}_{2,j} + \cdots + \mathbf{A}_{i,n}\mathbf{B}_{n,j}$
  - Defined only if the number of columns of the left matrix is the same as the number of rows of the right matrix
  - Note that $\mathbf{AB} \neq \mathbf{BA}$

$$\begin{bmatrix}\underline{2} & \underline{3} & \underline{4}\\ 1 & 0 & 0\end{bmatrix} \begin{bmatrix}0 & \underline{1000}\\ 1 & \underline{100}\\ 0 & \underline{10}\end{bmatrix} = \begin{bmatrix}3 & \underline{2340}\\ 0 & 1000\end{bmatrix}$$

Matrices

- ***Transpose*** of the matrix: $\mathbf{A}^T$ has the rows and columns exchanged

$$\left(\mathbf{A}^T\right)_{i,j} = \mathbf{A}_{j,i} \qquad \begin{bmatrix} 1 & 2 & 3 \\ 0 & -6 & 7 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} 1 & 0 \\ 2 & -6 \\ 3 & 7 \end{bmatrix}$$

  – Some properties

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \qquad \mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$$
$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \boldsymbol{B}^T \qquad \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$$
$$(\mathbf{A}^T)^T = \mathbf{A} \qquad (\mathbf{AB})^T = \boldsymbol{B}^T\mathbf{A}^T$$

- ***Square matrix***: has the same number of rows and columns

- ***Identity matrix*** ($\mathbf{I}_n$): has ones on the main diagonal, and zeros elsewhere

  – E.g.: identity matrix of size 3×3 :

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Matrices

- ***Determinant*** of a matrix, denoted by $\det(\mathbf{A})$ or $|\mathbf{A}|$, is a real-valued scalar encoding certain properties of the matrix
  - E.g., for a matrix of size 2×2: $$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad - bc$$
  - For larger-size matrices the determinant of a matrix id calculated as
$$\det(\mathbf{A}) = \sum_j a_{ij}(-1)^{i+j} det\left(\mathbf{A}_{(i,j)}\right)$$
  - In the above, $\mathbf{A}_{(i,j)}$ is a minor of the matrix obtained by removing the row and column associated with the indices i and j
- ***Trace*** of a matrix is the sum of all diagonal elements
$$\mathrm{Tr}(\mathbf{A}) = \sum_i a_{ii}$$
- A matrix for which $\mathbf{A} = \mathbf{A}^T$ is called a ***symmetric matrix***

Matrices

- Elementwise multiplication of two matrices **A** and **B** is called the ***Hadamard product*** or ***elementwise product***
  - The math notation is $\odot$

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \dots & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \dots & a_{2n}b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & a_{m2}b_{m2} & \dots & a_{mn}b_{mn} \end{bmatrix}$$

Matrix-Vector Products

- Consider a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$
- The matrix can be written in terms of its row vectors (e.g., $\mathbf{a}_1^T$ is the first row)

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_m^\top \end{bmatrix}$$

- The matrix-vector product is a column vector of length m, whose i^{th} element is the dot product $\mathbf{a}_i^T \mathbf{x}$

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_m^\top \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{x} \\ \mathbf{a}_2^\top \mathbf{x} \\ \vdots \\ \mathbf{a}_m^\top \mathbf{x} \end{bmatrix}$$

- Note the size: $\mathbf{A}(m \times n) \cdot \mathbf{x}(n \times 1) = \mathbf{Ax}(m \times 1)$

Matrix-Matrix Products

- To multiply two matrices $\mathbf{A} \in \mathbb{R}^{n\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times m}$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{km} \end{bmatrix}$$

- We can consider the matrix-matrix product as dot-products of rows in **A** and columns in **B**

$$\mathbf{C} = \mathbf{A}\mathbf{B} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_n^\top \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_m \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{b}_1 & \mathbf{a}_1^\top \mathbf{b}_2 & \cdots & \mathbf{a}_1^\top \mathbf{b}_m \\ \mathbf{a}_2^\top \mathbf{b}_1 & \mathbf{a}_2^\top \mathbf{b}_2 & \cdots & \mathbf{a}_2^\top \mathbf{b}_m \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_n^\top \mathbf{b}_1 & \mathbf{a}_n^\top \mathbf{b}_2 & \cdots & \mathbf{a}_n^\top \mathbf{b}_m \end{bmatrix}$$

- Size: $\mathbf{A}(n\times k) \cdot \mathbf{B}(k\times m) = \mathbf{C}(n\times m)$

Linear Dependence

- For the following matrix $\mathbf{B} = \begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix}$
- Notice that for the two columns $\mathbf{b}_1 = [2, 4]^T$ and $\mathbf{b}_2 = [-1, -2]^T$, we can write $\mathbf{b}_1 = -2 \cdot \mathbf{b}_2$
  - This means that the two columns are linearly dependent
- The weighted sum $a_1\mathbf{b}_1 + a_2\mathbf{b}_2$ is referred to as a linear combination of the vectors $\mathbf{b}_1$ and $\mathbf{b}_2$
  - In this case, a linear combination of the two vectors exist for which $\mathbf{b}_1 + 2 \cdot \mathbf{b}_2 = \mathbf{0}$
- A collection of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are ***linearly dependent*** if there exist coefficients $a_1, a_2, \ldots, a_k$ not all equal to zero, so that

$$\sum_{i=1}^{k} a_i \mathbf{v_i} = 0$$

- If there is no linear dependence, the vectors are ***linearly independent***

Matrix Rank

- For an $n \times m$ matrix, the ***rank*** of the matrix is the largest number of linearly independent columns
- The matrix $\mathbf{B}$ from the previous example has $rank(\mathbf{B}) = 1$, since the two columns are linearly dependent

$$\mathbf{B} = \begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix}$$

- The matrix $\mathbf{C}$ below has $rank(\mathbf{C}) = 2$, since it has two linearly independent columns
  - I.e., $\mathbf{c}_4 = -1 \cdot \mathbf{c}_1$, $\mathbf{c}_5 = -1 \cdot \mathbf{c}_3$, $\mathbf{c}_2 = 3 \cdot \mathbf{c}_1 + 3 \cdot \mathbf{c}_3$

$$\mathbf{C} = \begin{bmatrix} 1 & 3 & 0 & -1 & 0 \\ -1 & 0 & 1 & 1 & -1 \\ 0 & +3 & 1 & 0 & -1 \\ 2 & 3 & -1 & -2 & 1 \end{bmatrix}$$

Inverse of a Matrix

- For a square $n \times n$ matrix $\mathbf{A}$ with rank n, $\mathbf{A}^{-1}$ is its ***inverse matrix*** if their product is an identity matrix $\mathbf{I}$ $$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

- Properties of inverse matrices $$\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$$ $$\left(\mathbf{AB}\right)^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

$$\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})} adj(\boldsymbol{A})$$

- If $\det(A) = 0$ (i.e., $\text{rank}(A) < n$), then the inverse does not exist
  - A matrix that is not invertible is called a singular matrix

- Note that finding an inverse of a large matrix is computationally expensive
  - In addition, it can lead to numerical instability

- If the inverse of a matrix is equal to its transpose, the matrix is said to be orthogonal matrix $$\mathbf{A}^{-1} = \mathbf{A}^{T}$$

Pseudo-Inverse of a Matrix

- ***Pseudo-inverse*** of a matrix
  - Also known as Moore-Penrose pseudo-inverse

- For matrices that are not square, the inverse does not exist
  - Therefore, a pseudo-inverse is used

- If $m > n$, then the pseudo-inverse is $\mathbf{A}^{\dagger} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{T}$ and $\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{I}$
- If $m < n$, then the pseudo-inverse is $\mathbf{A}^{\dagger} = \mathbf{A}^{T}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)^{-1}$ and $\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{I}$
  - E.g., for a matrix with dimension $\mathbf{X}_{2\times3}$, a pseudo-inverse can be found of size $\mathbf{X}^{\dagger}_{3\times2}$, so that $\mathbf{X}_{2\times3}\mathbf{X}^{\dagger}_{3\times2} = \mathbf{I}_{2\times2}$

Tensors

- ***Tensors*** are n-dimensional arrays of scalars
  - Vectors are first-order tensors, $\mathbf{v} \in \mathbb{R}^n$
  - Matrices are second-order tensors, $\mathbf{A} \in \mathbb{R}^{m \times n}$
  - E.g., a fourth-order tensor is $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$

- Tensors are denoted with upper-case letters of a special font face (e.g., **X**, **Y**, **Z**)

- RGB images are third-order tensors, i.e., as they are 3-dimensional arrays
  - The 3 axes correspond to width, height, and channel
  - E.g., $224 \times 224 \times 3$
  - The channel axis corresponds to the color channels (red, green, and blue)

Eigen Decomposition

- ***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

$$\det(\mathbf{A} - \lambda\mathbf{I}\,) = 0$$

- If a matrix $\mathbf{A}$ has n linearly independent eigenvectors $\{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$ with corresponding eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, the eigen decomposition of $\mathbf{A}$ is given by

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$$

  - Columns of the matrix $\mathbf{V}$ are the eigenvectors, i.e., $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^n]$
  - $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues, i.e., $\boldsymbol{\Lambda} = [\lambda_1, \ldots, \lambda_n\,]$

- To find the inverse of the matrix A, we can use $\mathbf{A}^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^{-1}$
  - This involves simply finding the inverse $\boldsymbol{\Lambda}^{-1}$ of a diagonal matrix

Eigen Decomposition

- 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

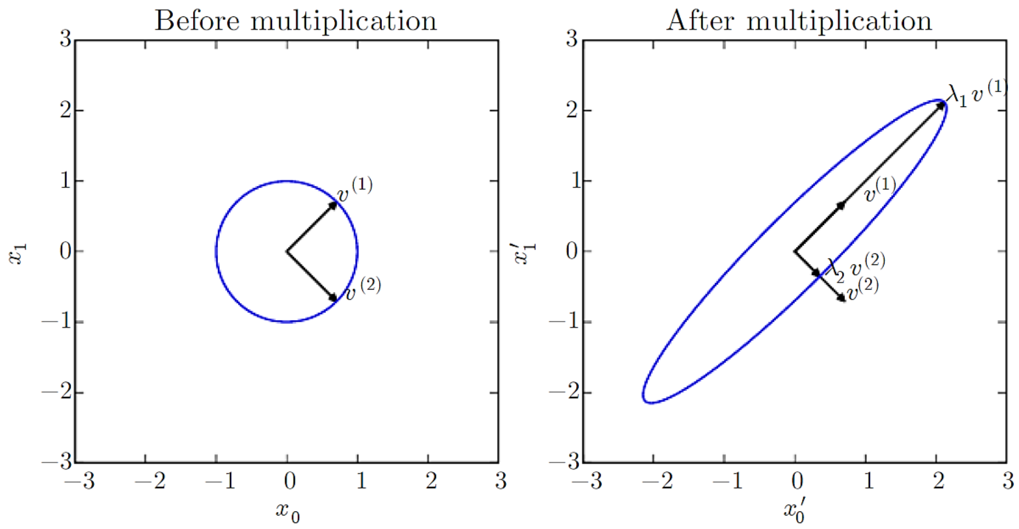
$$\det(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$$

  - 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 $\mathbf{A} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^{-1}$, where $\mathbf{V}$ is an orthogonal matrix

Eigen Decomposition

- Geometric interpretation of the eigenvalues and eigenvectors is that they allow to stretch the space in specific directions
  - Left figure: the two eigenvectors $\mathbf{v}^1$ and $\mathbf{v}^2$ are shown for a matrix, where the two vectors are unit vectors (i.e., they have a length of 1)
  - Right figure: the vectors $\mathbf{v}^1$ and $\mathbf{v}^2$ are multiplied with the eigenvalues λ_1 and λ_2
    - We can see how the space is scaled in the direction of the larger eigenvalue λ_1
- E.g., this is used for dimensionality reduction with PCA (principal component analysis) where the eigenvectors corresponding to the largest eigenvalues are used for extracting the most important data dimensions



Differential Calculus

Differential Calculus

- For a function $f: \mathbb{R} \rightarrow \mathbb{R}$, 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 $f'(c)$ is differentiable 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

- 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.** $\frac{d}{dx}\left(g(x) + h(x)\right) = \frac{dg}{dx}(x) + \frac{dh}{dx}(x)$.
- **Product rule.** $\frac{d}{dx}\left(g(x) \cdot h(x)\right) = g(x)\frac{dh}{dx}(x) + \frac{dg}{dx}(x)h(x)$.
- **Chain rule.** $\frac{d}{dx}g(h(x)) = \frac{dg}{dh}(h(x)) \cdot \frac{dh}{dx}(x)$.

Probability

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.

Random variables

- 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
- 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 $\mathrm{X} \sim P(X)$ 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

Axioms of probability

- 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, $P(\mathcal{A}_1 \cup \mathcal{A}_2) = P(\mathcal{A}_1)+P(\mathcal{A}_2)$

Multivariate Random Variables

- 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)

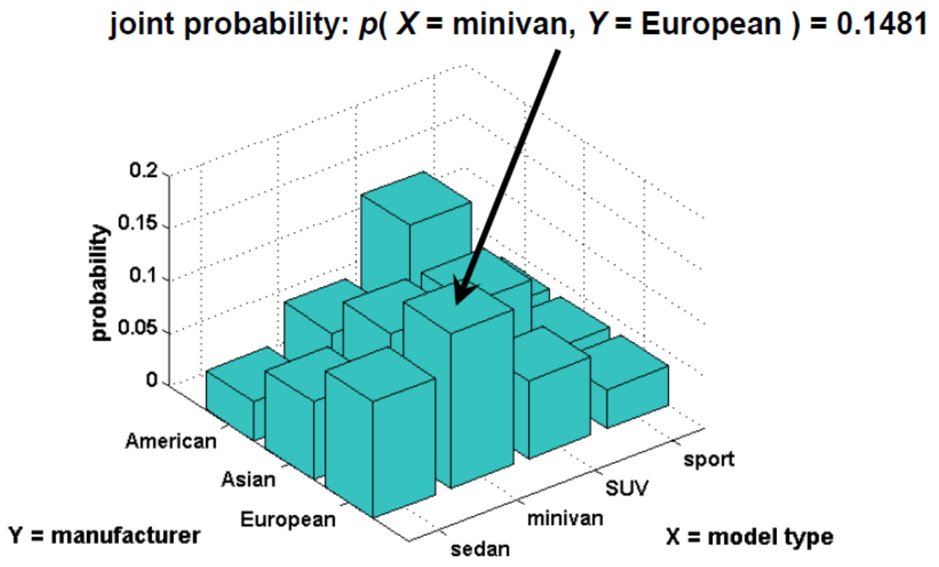
- Next, we will study probability distributions defined over multiple random variables
  - These include joint, conditional, and marginal probability distributions

- 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 $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$

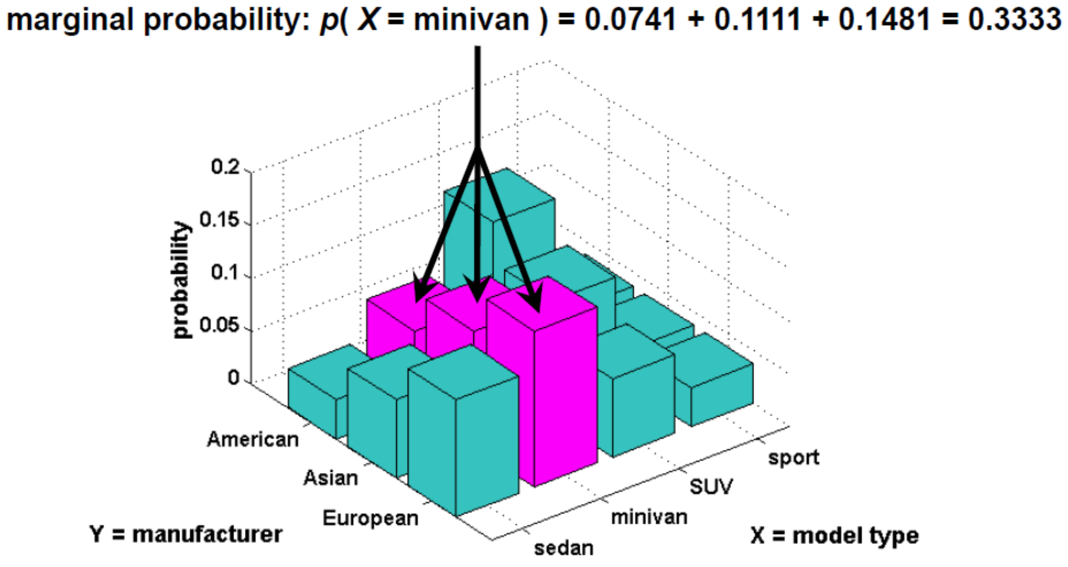
Joint Probability Distribution

- 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



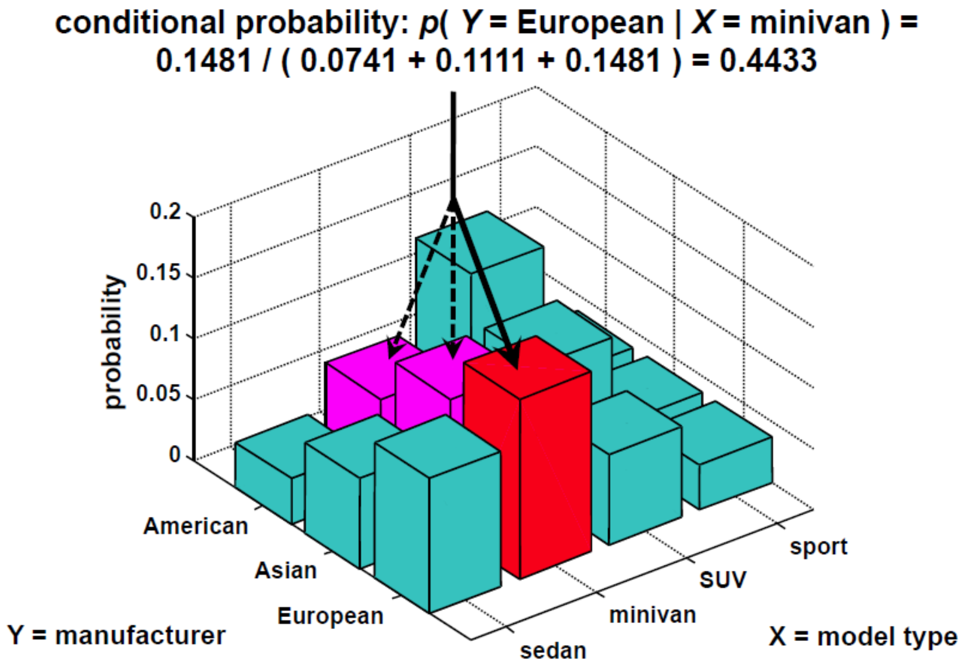
Marginal Probability Distribution

- ***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

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



Independence

- Two random variables X and Y are ***independent*** if the occurrence of Y does not reveal any information about the occurrence of X
  - E.g., two successive rolls of a die are independent

- Therefore, we can write: $P(X|\,Y) = P(X)$
  - The following notation is used: $X \perp Y$
  - Also note that for independent random variables: $P(X,Y) = P(X)P(Y)$

- Two random variables X and Y are ***conditionally independent*** given another random variable Z if and only if $P(X,Y|Z) = P(X|Z)P(Y|Z)$
  - This is denoted as $X \perp Y|Z$

Bayes' Theorem

- ***Bayes' theorem*** – allows to calculate conditional probabilities for one variable when conditional probabilities for another variable are known

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

- Also known as Bayes' rule
- Multiplication rule for the joint distribution is used: $P(X,Y) = P(Y|X)P(X)$

- The terms are referred to as:
  - $P(X)$, the prior probability, the initial degree of belief for X
  - $P(X|Y)$, the posterior probability, the degree of belief after incorporating the knowledge of Y
  - $P(Y|X)$, the likelihood of Y given X
  - P(Y), the evidence
  - Bayes' theorem: $\textbf{posterior probability} = \frac{\textbf{likelihood} \times \textbf{prior probability}}{\textbf{evidence}}$