CS 559: Machine Learning Fundamentals & Applications

Lecture 1: Mathematics for Machine Learning

Fall 2022

1870

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Outline

- Course Information
- Linear Algebra
- Analytic Geometry
- Vector Calculus
- Probability Theory

Course Information

Course Information

- Instructor: In Suk Jang, Ph.D.
- Course Web Address: https://sit.instructure.com/courses/61585
- Course Schedule: Online
- Contact Info: ijang@stevens.edu
- Virtual Office Hours: TBA
- Virtual Office Hour URL: https://stevens.zoom.us/j/5516841287
- Lecture day selection: Thursday 6:30-9 PM
- Course Syllabus: S22_CS559_Syllabus_Online.docx

Course Information

COURSE MATERIALS

- Bishop, Christopher M., 2006. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc. A comprehensive reference for bayesian theory that we will cover.
- Ian Goodfellow and Yoshua Bengio and Aaron Courville, 2016. Deep Learning, MIT Press. We will cover topics including basic neural networks, backpropagation, and CNN.
- Hastie, Trevor, and Tibshirani, Robert and Friedman, Jerome, 2008. The Elements of Statistical Learning. Second Edition, Springer New York Inc.
- **The main lecture will be following Bishop**. However, students are not required to purchase the book.

Course Information

COURSE REQUIREMENTS

- **Quiz (10%):** We will have a short online quiz, about 10 to 15 minutes long, for each topic. There are a total of 7 quizzes. Each quiz will be available for a week. The topic will come from the lecture and questions will be conceptual.
- **Homework (25%):** There will be a total of three bi-weekly assignments. Each assignment is centered around an application and will also deepen your understanding of the theoretical concepts. Every homework will be available from Monday at 11:59 PM for two weeks and must be submitted in two weeks.
- **Project 1 (25%):** The first project focuses on data pre-processing practices and applications of some supervised learning techniques.
- **Project 2 (25%):** The second project focuses on data applications of some supervised learning and deep learning techniques.
- **Final Exams (15%):** The final exam is to evaluate your understanding of the whole course.

LATE SUBMISSION POLICY

- Applies 15% reduction for every 24 hours.

Course Information

Work Environment

- Language: Python
- Work Platform: Visual Studio, Visual Studio Code, Jupyter Notebook, Google Code

1.1. Linear Algebra

1.1.1. Basic Matrix Identities

A matrix **A** has elements A_{ij} where i indexes the rows, and j indexes the columns.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \qquad A_{ij} \in \mathbb{R}.$$

1.1.2. Matrix Addition and Multiplication

The sum of two matrices $\boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$ is the element-wise sum,

$$\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

The product of two matrices $\boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{B} \in \mathbb{R}^{n\times k}$ is the pair-wise sum,

$$\boldsymbol{C} = \boldsymbol{AB} \in \mathbb{R}^{m\times k}, c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj}\,, i = 1, \dots, m, j = 1, \dots, k.$$

1.1.2. Matrix Addition and Multiplication

Matrix Multiplication Properties:

- Not commutative: $\boldsymbol{AB} \neq \boldsymbol{BA}$
- Associative: $\forall \boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{B} \in \mathbb{R}^{n\times p}, \boldsymbol{C} \in \mathbb{R}^{p\times q}: (\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$
  - $(\lambda\psi)\boldsymbol{C} = \lambda(\psi)\boldsymbol{C}$ where λ and ψ are constants.
  - $\psi(\boldsymbol{BC}) = (\psi\boldsymbol{B})\boldsymbol{C} = \boldsymbol{B}(\psi\boldsymbol{C}) = (\boldsymbol{BC})\psi$, $\boldsymbol{B} \in \mathbb{R}^{m\times n}, \boldsymbol{C} \in \mathbb{R}^{n\times k}$
- Distributive: $\forall \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}, \boldsymbol{C} \in \mathbb{R}^{n\times p}:\ (\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$
  - $(\lambda + \psi)\boldsymbol{C} = \lambda\boldsymbol{C} + \psi\boldsymbol{C}$

1.1.3. Inverse and Transpose

A matrix $\boldsymbol{I}_N$ is the $N \times N$ *identity* matrix (also called the unit matrix)

$$I_N = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}.$$

The inverse of $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{n}}$, $\boldsymbol{A}^{-1}$, satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-\mathbf{1}} = \boldsymbol{A}^{-\mathbf{1}}\boldsymbol{A} = \boldsymbol{I}.$$

If the inverse exists, $\boldsymbol{A}$, is called *regular/invertible/nonsingular.*

Consider two matrices

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \text{ and } \boldsymbol{A}' = \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix},$$

the product of $\boldsymbol{A}\boldsymbol{A}'$ is then

$$\boldsymbol{A}\boldsymbol{A}' = \begin{bmatrix} a_{11}a_{22} - a_{12}a_{21} & 0 \\ 0 & a_{11}a_{22} - a_{12}a_{21} \end{bmatrix} = (a_{11}a_{22} - a_{12}a_{21})\boldsymbol{I}.$$

If and only if $a_{11}a_{22} - a_{12}a_{21} \neq 0$,

$$\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

1.1.3. Inverse and Transpose

Inverse Matrix Properties:

- $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I} = \boldsymbol{A}^{-1}\boldsymbol{A}$
- $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$
- $(\boldsymbol{A} + \boldsymbol{B})^{-1} \neq \boldsymbol{A}^{-1} + \boldsymbol{B}^{-1}$

The *Woodbury identity* is

$$\left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} + \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}.$$

The transpose matrix $\boldsymbol{A}^T$ has elements $(A^T)_{ij} = A_{ji}$. From the definition of transpose, a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is *symmetric* if $\boldsymbol{A} = \boldsymbol{A}^T$.

Transpose Properties:

- $(\boldsymbol{A}^T)^T = \boldsymbol{A}$
- $(\boldsymbol{A} + \boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$
- $(\boldsymbol{A}\boldsymbol{B})^T = \boldsymbol{B}^T\boldsymbol{A}^T$

If $\boldsymbol{A}$ is invertible, then $(\boldsymbol{A}^{-1})^T = (\boldsymbol{A}^T)^{-1} = \boldsymbol{A}^{-T}$.

1.1.4. Vector Spaces

Vector: Each row or column in a matrix $\boldsymbol{A}$ is called a *vector*. A real-valued vector $V = (\mathcal{V}, +, \cdot)$ is a set $\mathcal{V}$ with two operations

$$+: \mathcal{V} \times \mathcal{V} \rightarrow \mathcal{V}$$
$$\cdot: \mathbb{R} \times \mathcal{V} \rightarrow \mathcal{V}$$

where $+$ is the vector addition and $\cdot$ is a multiplication by a scalar.

Vector Subspace: Suppose V is a vector space and $\mathcal{U} \subseteq \mathcal{V}, \mathcal{U} \neq \emptyset$. Then U is a vector subspace of V if U is a **vector subspace** with the vector space operations restricted to $\mathcal{U} \times \mathcal{U}$ and $\mathbb{R} \times \mathcal{U}$.

Linear Combination: A vector space V and a finite number of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in V$. Then, every $\boldsymbol{v} \in V$ of the form

$$\boldsymbol{v} = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i \in V$$

with $\lambda_i \in \mathbb{R}$ is a *linear combination* of the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$.

1.1.5. Basis and Rank

Span: When a set of vectors $\mathcal{A} = \{x_1, \dots, x_k\}$ is in a vector space V and if every vector can be expressible in a linear combination format, $\mathcal{A}$ is called the *generating set* of V. The set of all linear combinations of vectors in $\mathcal{A}$ is called the **span** of $\mathcal{A}$. The denotation $V = \text{span}[\mathcal{A}]$ means $\mathcal{A}$ spans the vector space V.

Basis: If there exists no smaller set $\mathcal{A} \subseteq V$ that spans, every linearly independent generating set of V is called a **basis** of V**.**

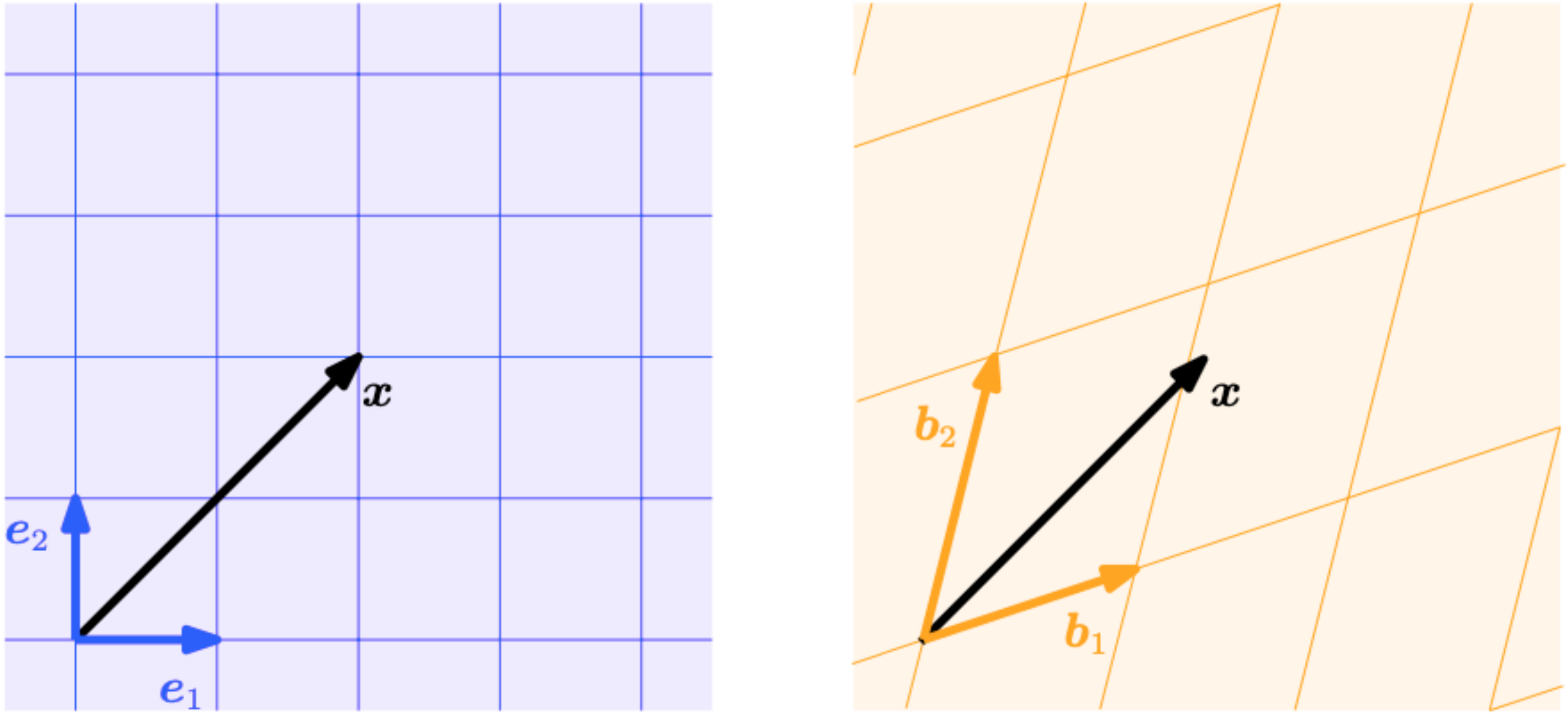
Rank: The number of linear independent columns of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ equals the number of linearly independent rows and is called the **rank** of $\boldsymbol{A}$.

1.1.6. Linear Mapping

Linear Mapping: For vector spaces V, W, a mapping $\Phi: V \to W$ is called a **linear mapping** if

$$\forall \boldsymbol{x}, \boldsymbol{y} \in V, \forall \lambda, \psi \in \mathbb{R}: \Phi(\lambda \boldsymbol{x} + \psi \boldsymbol{y}) = \lambda \Phi(\boldsymbol{x}) + \psi \Phi(\boldsymbol{y}).$$

- If $\Phi: V \to W, \Psi: \mathrm{V} \to \mathrm{W}$ are linear, then $\Phi + \Psi$ and $\lambda\Phi, \lambda \in \mathbb{R}$, are linear, too.



1.1.6. Linear Mapping

In a vector space V, an ordered basis $B = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)$ of V, a unique linear combination of $\boldsymbol{x} \in V$ is formulated:

$$\boldsymbol{x} = \alpha_1 \boldsymbol{b_1} + \cdots + \alpha_n \boldsymbol{b_n}$$

where $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_n] \in \mathbb{R}^n$ are coordinates of $\boldsymbol{x}$ w.r.t. B.

Consider vector spaces V, W with corresponding bases $B = (b_1, \dots, b_n)$ and $C = (c_1, \dots, c_m)$. The linear mapping $\Phi: V \rightarrow W$ for $j \in \{1, \dots, n\}$ is

$$\Phi(\boldsymbol{b}_j) = \sum_{i=1}^{m} \alpha_{ij} \boldsymbol{c}_i \, .$$

The $m \times n$ matrix $\boldsymbol{A_\Phi}$ is a transformation matrix of Φ.

1.2. Analytic Geometry

1.2.1. Norms

Norm: A **norm** on a vector space V is a function

$$||\cdot||: V \to \mathbb{R},$$
$$\boldsymbol{x} \mapsto ||\boldsymbol{x}||,$$

which assigns each vector $\boldsymbol{x}$ its *length* $||\boldsymbol{x}|| \in \mathbb{R}$, such that for all $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in V$ the following hold:

- *Absolutely homogeneous*: $||\lambda \boldsymbol{x}|| = |\lambda| ||\boldsymbol{x}||$
- *Triangle inequality*: $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$
- *Positive definite*: $||x|| \geq 0$ and $||\boldsymbol{x}|| = 0 \Leftrightarrow \boldsymbol{x} = 0$.

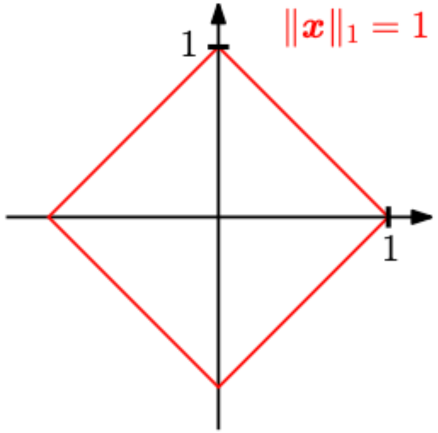
1.2.1. Norms

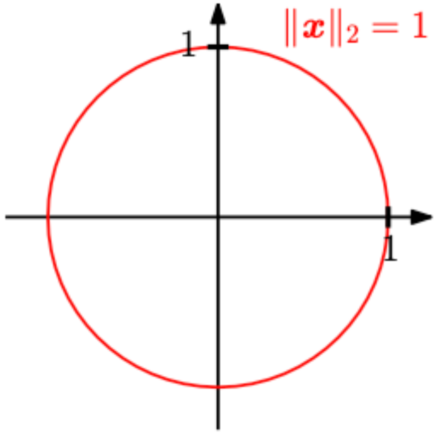
Manhattan Norm: The **Manhattan norm** (also called l_1) on $\mathbb{R}^n$ is defined for $\boldsymbol{x} \in \mathbb{R}^n$ as

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

Euclidean Norm: The **Euclidean norm (**also called l_2**)** of $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

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





1.2.2. Inner Product

Dot Product: The dot product takes two equal-length sequences of numbers or vectors and returns a single number as follow

$$\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{x}^T \cdot \mathrm{y} = \sum_{i=1}^{n} x_i y_i \,.$$

A *bilinear mapping* Ω is a mapping with two arguments, and it is linear in each argument that holds for all $\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z} \in V$, λ, ψ ∈ ℝ such that

$$\Omega(\lambda\boldsymbol{x} + \psi\boldsymbol{y}, \boldsymbol{z}) = \lambda\Omega(\boldsymbol{x}, \boldsymbol{z}) + \psi\Omega(\boldsymbol{y}, \boldsymbol{z})$$
$$\Omega(x, \lambda y + \psi z) = \lambda\Omega(x, y) + \psi\Omega(x, z)$$

- Ω is called *symmetric* if $\Omega(x, y) = \Omega(y, x)$ for all $x, y \in V$, i.e., the order of the argument does not matter.
- Ω is called *positive definite* if

$$\forall \boldsymbol{x} \in V\backslash\{\boldsymbol{0}\} : \Omega(\boldsymbol{x}, \boldsymbol{x}) > 0, \Omega(\boldsymbol{0}, \boldsymbol{0}) = 0.$$

- A positive definite, symmetric bilinear mapping $\Omega: V{\times}V \rightarrow \mathbb{R}$ is an *inner product* on V.
- The denotation of inner product is $\langle x, y\rangle$.
- An inner product is not always the dot product! For example, we define the inner product in $\mathbb{R}^2$ is

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \ x_1y_1 - (x_1y_2 + x_2y_1) + 2x_2y_2,$$

  it is different from $\boldsymbol{x}^T\boldsymbol{y} = x_1y_1 + x_2y_2$.

- The inner product of two vectors

$$\sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y}\rangle} = ||\boldsymbol{x} - \boldsymbol{y}||$$

  is the *distance* between two vectors.

1.2.2. Inner Product

Consider an n-dimensional vector space V with an inner product and an ordered basis $B = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)$ of V. Any vectors $\boldsymbol{x}, \boldsymbol{y} \in V$ form as linear combinations $\boldsymbol{x} = \sum_{i=1}^{n} \psi_i \boldsymbol{b}_i \in V$ and $\boldsymbol{y} = \sum_{j=1}^{n} \lambda_i \boldsymbol{b}_j \in V$ where $\forall \lambda_j, \psi_i \in \mathbb{R}$. The inner product is

$$\forall \boldsymbol{x}, \boldsymbol{y}\colon \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \left\langle \sum_{i=1}^{n} \psi_i \boldsymbol{b}_i, \sum_{j=1}^{n} \lambda_j \boldsymbol{b}_j \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_i \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle \lambda_j = \hat{\boldsymbol{x}}^{\boldsymbol{T}} \boldsymbol{A} \hat{\boldsymbol{y}},$$

where $A_{ij} = \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle$ and $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$ are unit vectors $\left(\hat{\boldsymbol{x}} = \frac{\boldsymbol{x}}{||\boldsymbol{x}||}, \hat{\boldsymbol{y}} = \frac{\boldsymbol{y}}{||\boldsymbol{y}||} \right)$ with respect to B.

A symmetric matrix that satisfies

$$\forall \boldsymbol{x} \in V \backslash \{\boldsymbol{0}\}\colon \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0$$

is called *symmetric, positive semidefinite.*

1.2.3. Orthogonality

Orthogonality: Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are **orthogonal** if and only if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$ and its denotation is $\boldsymbol{x} \perp \boldsymbol{y}$. That is the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$ is 0,

$$\cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle}} = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\sqrt{\boldsymbol{x}^T \boldsymbol{x} \boldsymbol{y}^T \boldsymbol{y}}} = 0.$$

A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an *orthogonal matrix* i.f.f. its columns are orthogonal so that

$$\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I} = \boldsymbol{A}^T \boldsymbol{A},$$

which implies that

$$\boldsymbol{A}^{-\mathbf{1}} = \boldsymbol{A}^T.$$

Orthonormal Basis: The basis is called an **orthonormal basis** if an n-dimensional vector and a basis $\{\boldsymbol{b}_{\mathbf{1}}, \dots, \boldsymbol{b}_{\boldsymbol{n}}\}$ hold

$$\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = 0 \text{ for } i \neq j \text{ and } \langle \boldsymbol{b}_i, \boldsymbol{b}_i \rangle = 1$$

for all $i, j = 1, \dots, n$.

Projection: A linear mapping $\pi: V \to U$ is called a **projection** if $\pi^2 = \pi \circ \pi = \pi$ where $U \subseteq V$. The *projection matrices* $\boldsymbol{P}_{\boldsymbol{\pi}}$ exhibit the property that $\boldsymbol{P}_\pi^2 = \boldsymbol{P}_\pi$.

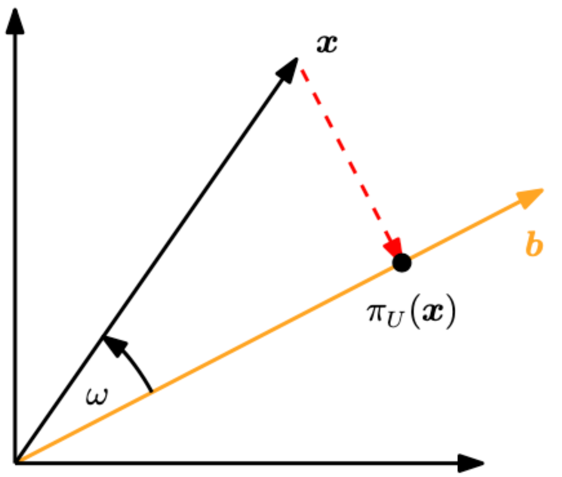
1.2.3. Orthogonality

- The segment $\pi_U(\boldsymbol{x}) - \boldsymbol{x}$ is orthogonal to U and therefore the basis vector $\boldsymbol{b}$ of U, $\langle \pi_U(\boldsymbol{x}) - \boldsymbol{x}, \boldsymbol{b} \rangle = 0$.
- The projection $\pi_U(\boldsymbol{x})$ of $\boldsymbol{x}$ onto U must be an element of U and, therefore, $\boldsymbol{b}$ spans U, $\pi_U(\boldsymbol{x}) = \lambda \boldsymbol{b}$ where λ is the coordinate

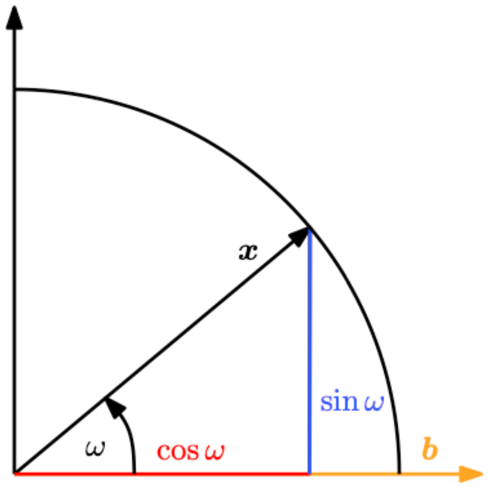
$$\lambda = \frac{\boldsymbol{b}^T \boldsymbol{x}}{\boldsymbol{b}^T \boldsymbol{b}} = \frac{\boldsymbol{b}^T \boldsymbol{x}}{||\boldsymbol{b}||^2}$$

and

$$\pi_U(\boldsymbol{x}) = \lambda \boldsymbol{b} = \frac{\boldsymbol{b}^T \boldsymbol{x}}{||\boldsymbol{b}||^2} \boldsymbol{b}.$$



(a) Projection of $\boldsymbol{x} \in \mathbb{R}^2$ onto a subspace U with basis vector $\boldsymbol{b}$.



(b) Projection of a two-dimensional vector $\boldsymbol{x}$ with $\|\boldsymbol{x}\| = 1$ onto a one-dimensional subspace spanned by $\boldsymbol{b}$.

1.2.4. Determinant and Trace

Determinant: A **determinant** is a mathematical object in the analysis and solution of systems of linear equations. It is only defined for *square matrices* and it maps a square matrix onto a real number. Recall the inverse matrix of $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$, $\boldsymbol{A}^{-1}$,

$$\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

can be expressed using the determination as following

$$\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

where

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

If $n = 3$,

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}(a_{22}a_{33} - a_{23}a_{32})$$
$$-a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

1.2.4. Determinant and Trace

For any $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\det(\boldsymbol{A})$ also can be computed as

- Expansion along column *j*:

$$\det(\boldsymbol{A}) = \sum_{k=1}^{n} (-1)^{k+j} a_{kj} \det\left(\boldsymbol{A}_{kj}\right).$$

- Expansion along row *j*:

$$\det(\mathbf{A}) = \sum_{\mathrm{k}=1}^{\mathrm{n}} (-1)^{k+j} a_{jk} \det\left(\boldsymbol{A}_{jk}\right).$$

where $\boldsymbol{A}_{kj} \in \mathbb{R}^{(n-1)\times(n-1)}$ is the *submatrix* of $\boldsymbol{A}$ that is obtained when row k and column j are deleted.

For example,

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = (-1)^{1+1} a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + (-1)^{1+2} a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + (-1)^{1+3} a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

The determinant exhibits the following properties:

- $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$
- If $\mathbf{A}$ is invertible, then $\det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})}$
- Adding a multiple of a column/row to another one does not change $\det(\boldsymbol{A})$.
- Multiplication of a column/row with $\lambda \in \mathbb{R}$ scales $\det(\mathbf{A})$ by λ: $\det(\mathbf{A}) = \lambda^{\mathrm{n}} \det(\boldsymbol{A})$.
- Swapping two rows/columns changes the sign of $\det(\boldsymbol{A})$.

1.2.4. Determinant and Trace

Trace: The **trace** of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is defined as

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}$$

and satisfies the following properties:

- $\operatorname{tr}(\boldsymbol{A}+\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B})$ for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$.
- $\operatorname{tr}(\alpha\boldsymbol{A}) = \alpha\operatorname{tr}(\boldsymbol{A}), \alpha \in \mathbb{R}$ for $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.
- $\operatorname{tr}(\boldsymbol{I}_n) = n$.
- $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$ for $\boldsymbol{A} \in \mathbb{R}^{n\times k}, \boldsymbol{B} \in \mathbb{R}^{k\times n}$.
- For vectors, $\boldsymbol{x}, \boldsymbol{y}$, $\operatorname{tr}\left(\boldsymbol{x}\boldsymbol{y}^{\boldsymbol{T}}\right) = \boldsymbol{y}^{T}\boldsymbol{x} \in \mathbb{R}$.

1.2.5. Eigenvalues and Eigenvectors

Eigenvalues and Eigenvector: Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an **eigenvalue** of $\boldsymbol{A}$ and $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ is the corresponding **eigenvector** of $\boldsymbol{A}$ if

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

where λ is eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. It is equivalent to

$$(\boldsymbol{A} - \lambda \boldsymbol{I}_n) = 0$$

and therefore

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}_n) = 0.$$

1.3. Vector Calculus

1.3.1. Differentiation of Univariate Functions

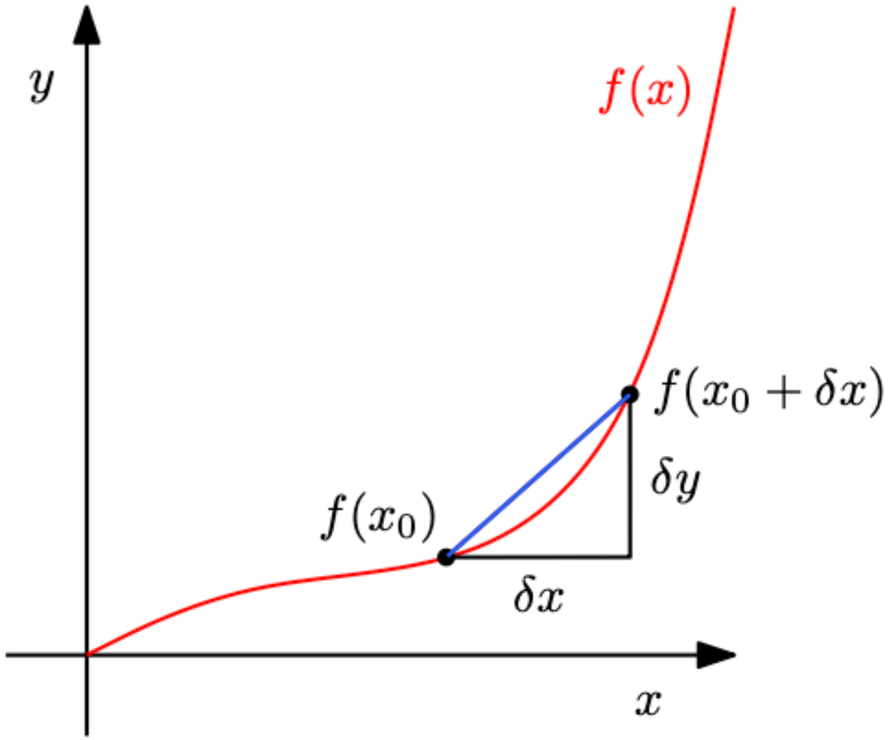
Derivative: The **derivative** of a function f at x is defined as the limit

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

and it points in the direction of steepest ascent of f.

Differentiation rules are

- Product rule: $\big(f(x)g(x)\big)' = f'(x)g(x) + f(x)g'(x)$
- Quotient rule: $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g(x)}{\big(g(x)\big)^2}$
- Sum rule: $\big(f(x) + g(x)\big)' = f'(x) + g'(x)$
- Chain rule:



1.3.2. Partial Differentiation and Gradients

Partial derivatives and Gradient: For a function $f: \mathbb{R}^n \to \mathbb{R}, \boldsymbol{x} \mapsto f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$ of n variables $x_1, \dots, x_n$, the **partial derivatives** are defined as

$$\frac{\partial f}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \dots, x_n) - f(\boldsymbol{x})}{h}$$
$$\vdots$$
$$\frac{\partial f}{\partial x_n} = \lim_{h \to 0} \frac{f(x_1, x_2, \dots, x_n + h) - f(\boldsymbol{x})}{h}$$

and collect them in the row vector

$$\nabla_x f = \frac{df}{d\boldsymbol{x}} = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \cdots, \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right] \in \mathbb{R}^{1 \times n}$$

1.3.3. Matrix Derivatives

The derivative of a vector **a** w.r.t. a scalar x is itself a vector whose components are given by

$$\left(\frac{\partial \boldsymbol{a}}{\partial x}\right)_i = \frac{\partial a_i}{\partial x}.$$

Similarly,

$$\left(\frac{\partial x}{\partial \boldsymbol{a}}\right)_i = \frac{\partial x}{\partial a_i} \text{ and } \left(\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right)_{ij} = \frac{\partial a_i}{\partial b_j}.$$

Furthermore,

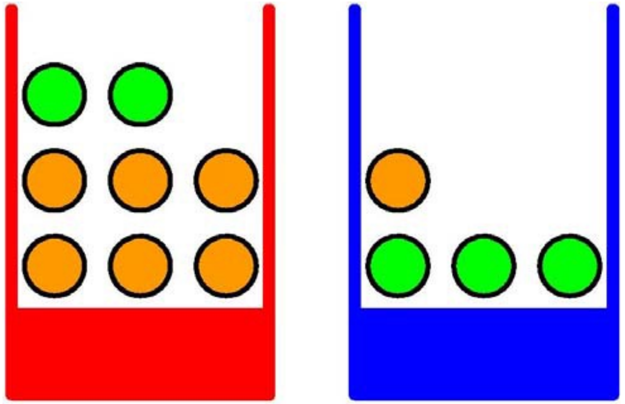
$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^T\boldsymbol{a}) = \frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{a}^T\boldsymbol{x}) = \boldsymbol{a}$$

$$\frac{\partial}{\partial x}(\boldsymbol{AB}) = \frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{B} + \boldsymbol{A}\frac{\partial \boldsymbol{B}}{\partial x}$$

$$\frac{\partial}{\partial x}\left(\boldsymbol{A}^{-\mathbf{1}}\right) = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$$

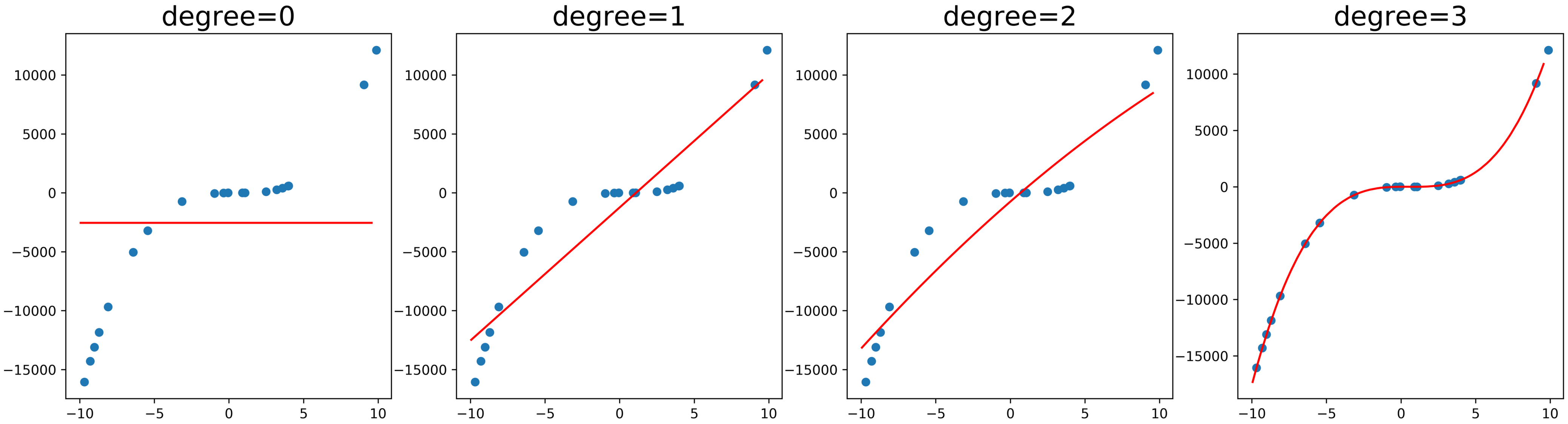
1.4. Probability Theory

1.4.0. – Motivation: Probability Theory



- Let us look at the following example:
  - We have two boxes, one red and one blue
  - Red box: 2 apples and 6 oranges
  - Blue box: 3 apples and 1 orange
  - Pick red box 40% of the time and blue box 60% of the time, then pick one item of fruit
  - Question1: what is the overall probability that the selection procedure will pick an apple?
  - Question2: given that we have chosen an orange, what is the probability that the box we chose was the blue one?

1.4.0. Motivation: Curve Fit



1.4.1. Probability and Random Variables

- **Sample Space Ω:** the set of all possible outcomes of the experiment.
- **Event Space**: the space of potential results of the experiment. A subset of sample space is in the event space.
- **Probability**: measurements the probability or degree of belief that the event will occur.

1.4.2. Discrete Probabilities

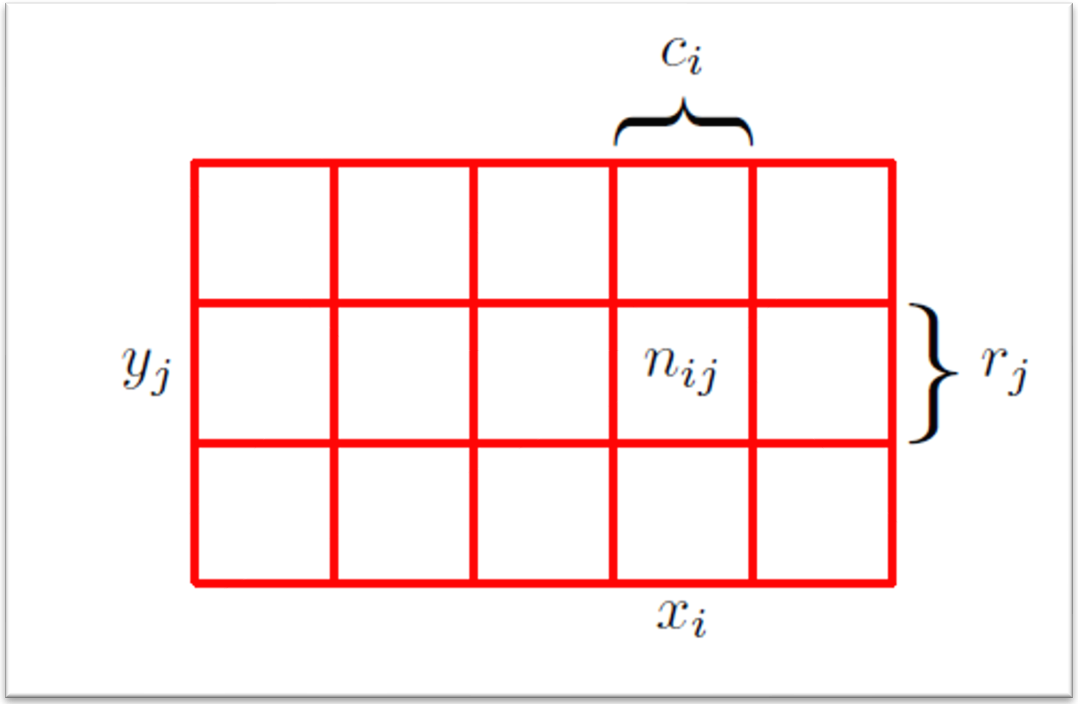
When the target space is discrete, the probability distribution of multiple random variables as filling out a (multidimensional) array of numbers.

Probability mass function: If x is a discrete variable, then $p(x)$ is sometimes called a *probability mass function* which implies as a set of probability masses concentrated at the allowed values of x.

1.4.2. Discrete Probabilities

Consider

- X can take any of the values x_i where $i = 1, \dots, M$.
- Y can take the values y_j where $j = 1, \dots, L$.
- A total of N trials in both of the variables X and Y.
- Let the number of trials for $X = x_i, Y = y_j$ be n_{ij}.
- Let the number of trials in which X takes the value x_i (irrespective of the value that Y takes) be denoted by c_i, and let the number of trials in which Y takes the value y_j be r_j.



1.4.2. Discrete Probabilities

- The probability that X will take the value of x_i and Y will take the value y_j, $p(X = x_i, Y = y_j)$, is called the *joint* probability of $X = x_i$ and $Y = y_j$.

$$p\left(X = x_i, Y = y_j\right) = \frac{n_{ij}}{N}.$$

- The probability that X takes the value x_i irrespective of the value of Y is

$$p(X = x_i) = \frac{c_i}{N}.$$

- **Sum Rule**: We have $c_i = \sum_j n_{ij}$ and therefore,

$$p(X = x_i) = \sum_{j=1}^{L} p\left(X = x_i, Y = y_j\right).$$

- Similarly, the probability that Y takes the value y_j irrespective of the value of X is

$$p\left(Y = y_j\right) = \frac{r_j}{N} = \sum_{i=1}^{M} p\left(X = x_i, Y = y_j\right).$$

- $p(X = x_i)$ and $p\left(Y = y_j\right)$ are sometimes called the *marginal* probability.

1.4.2. Discrete Probabilities

Consider for which $X = x_i$, the fraction of instances for which $Y = y_j$ is

$$p(Y = y_j | X = x_i)$$

called the *conditional* probability of $Y = y_j$ given $X = x_i$. It is obtained by finding the fraction of those points in i fall in (i, j):

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}.$$

The conditional probability of X given Y is

$$p(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}.$$

1.4.3. Continuous Probability

- Consider the target space with real continuous numbers $\mathbb{R}$.
- Most often, we pretend that we perform operations as we have discrete probability spaces with finite spaces. However, the simplification is not precise when
  - if operations were infinitely repeated. or
  - if a single point were drawn from an interval.

- **Probability Density Function** (pdf): If the probability of a real-valued variable x falling in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$, then $p(x)$ is called the *probability density* over x.

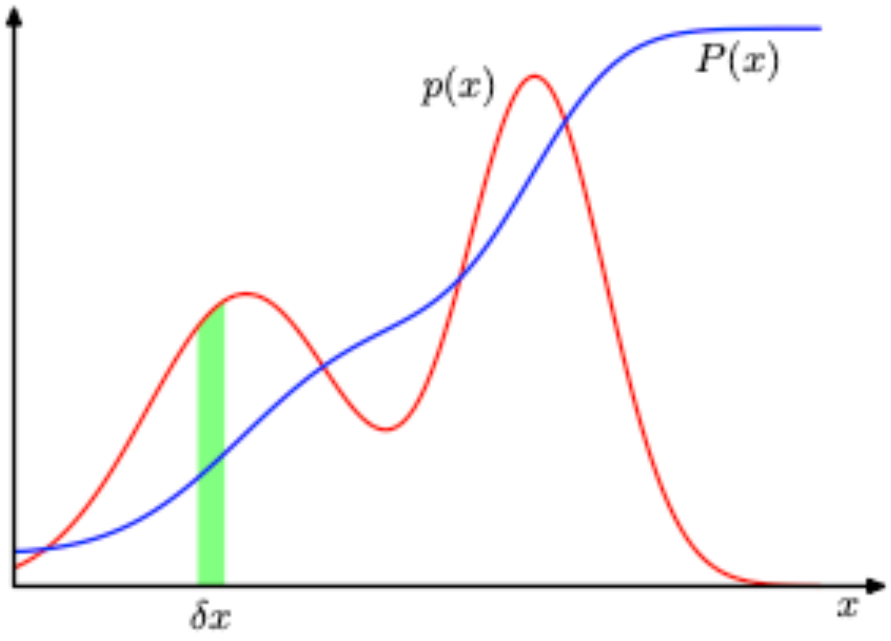
- The probability that x will lie in an interval (a, b) is then given by

$$p\big(x \in (a, b)\big) = \int_a^b p(x)\, dx\,.$$

- The probability density $p(x)$ must satisfy the two conditions:

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)dx = 1.$$



1.4.3. Continuous Probability

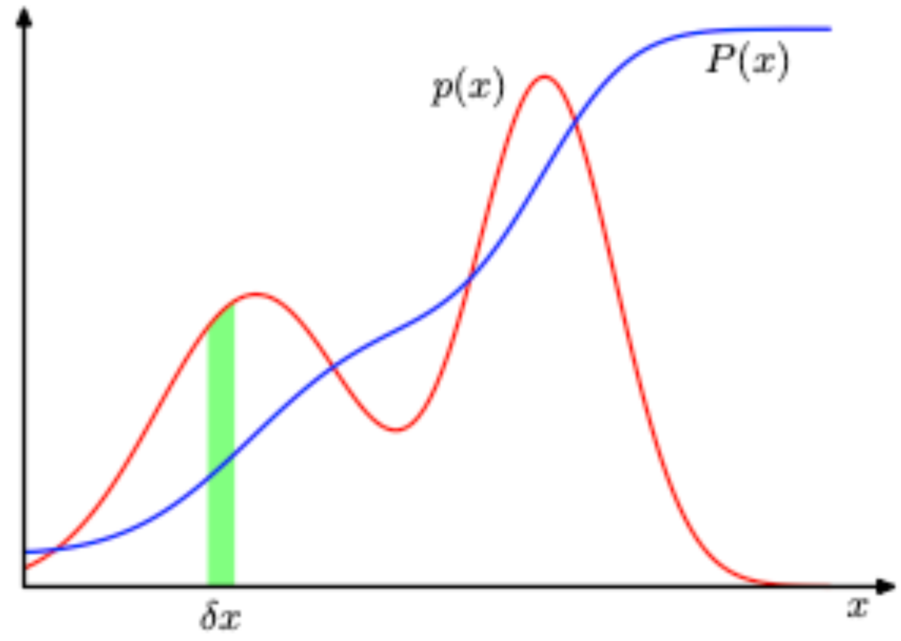
- If we consider a change variable $x = g(y)$, then a function $f(x)$ becomes $f\big(g(y)\big)$. This means that a probability density $p_x(x)$ that corresponds to a density $p_y(y)$ with respect to the new variable y, where the suffixes denote the fact that $p_x(x)$ and $p_y(y)$ are different densities. The observation then transforms into the range $(y, y + \delta y)$ where $p_x(x)\delta x \sim p_y(y)\delta y$ and

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right| = p_x\big(g(y)\big)|g'(y)|.$$

- **Cumulative distribution function (cdf):** The probability that x lies in the interval $(-\infty, z)$ is given by the *cumulative distribution function* defined by

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

- Which satisfies $P'(z) = p(x)$.



1.4.4. Sum Rule, Product Rule, and Bayes' Theorem

Sum Rule: The sum ule is the *marginalization property*. It relates the joint distribution to a marginal distribution via

$$p(\boldsymbol{x}) = \begin{cases} \sum_{y \in Y} p(\boldsymbol{x}, \boldsymbol{y}), & \text{if } \boldsymbol{y} \text{ is discrete} \\ \int_Y p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y}, & \text{if } \boldsymbol{y} \text{ is continuous} \end{cases}.$$

Product Rule: It relates the joint distribution to the conditional distribution via

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}).$$

- Every joint distribution of two random variables can be factorized of two other distributions.
- The ordering of random variables is arbitrary, $p(x, y) = p(x|y)p(y)$.

1.4.4. Sum Rule, Product Rule, and Bayes' Theorem

- In machine learning and Bayesian statistics, we are often interested in making inferences of unobserved random variables given that we have observed other random variables.
-
- Assume some prior knowledge $p(Y)$ about an observed random variable Y and some relationship $p(X|Y)$ between X and a second random variable Y, which we can observe.

- **Bayes' Theorem (rule or law)**: From the product rule, together with the symmetry property $p(X, Y) = p(Y, X)$, we can build the following relationship called *Bayes' theorem* between conditional probabilities

$$\text{Posterior} = \text{Likelihood} \cdot \frac{\text{Prior}}{\text{Event}} \rightarrow p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}.$$

- We can view the denominator as the normalizer

$$p(X) = \sum_{y} p(X|Y)p(Y)$$

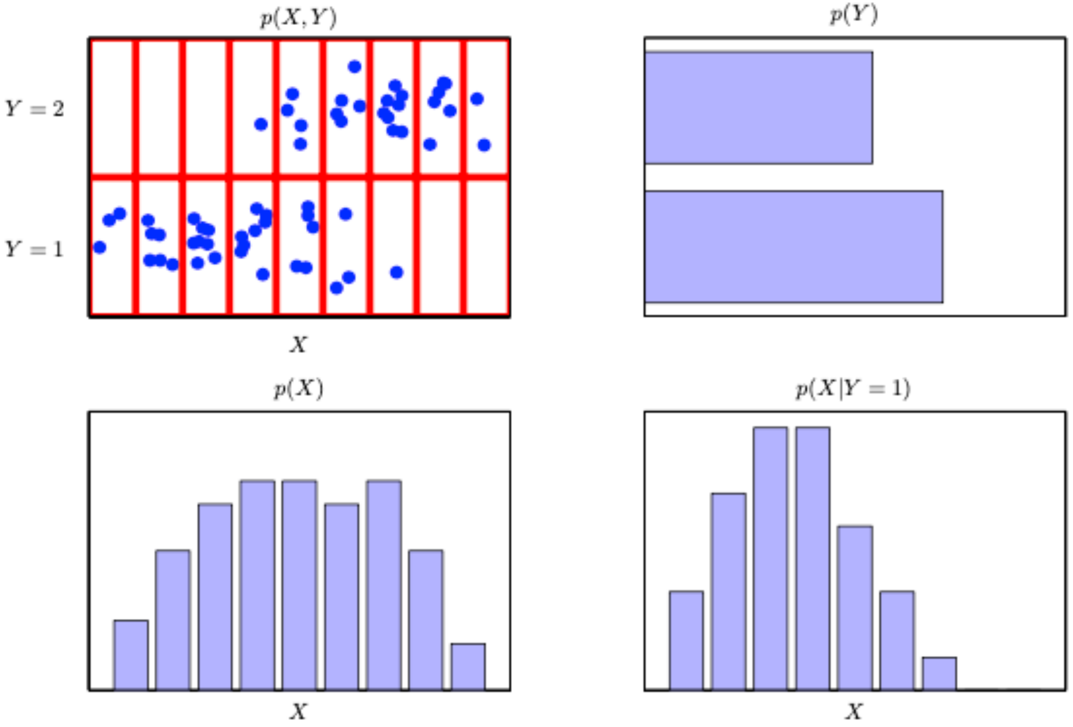
using the sum rule.

1.4.4. Sum Rule, Product Rule, and Bayes' Theorem

- The *prior* $P(X)$ probability encapsulates the subjective prior knowledge of the unobserved variable X before observing any data.

- The *likelihood* $p(X|Y)$ describes how X and Y are related. It is the probability of the data X if the latent variable Y were to know. The likelihood is not a distribution in Y, but only in X.

- The *posterior* $P(Y|X)$ is the quantity of interest in Bayesian statistics.



1.4.4. Sum Rule, Product Rule, and Bayes' Theorem

- From a Bayesian perspective, probability provides a quantification of uncertainty.

- Considering the curve-fit example, we can use the probability theory to describe the uncertainty in model parameters and the models.

- Let the function be $t(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$.
  - Assume we know the prior probability of $\boldsymbol{w} = [w_0, w_1, w_2, w_3], p(\boldsymbol{w})$.
  - The effect of the observed data $\mathcal{D} = \{t_1, \dots, t_N\}$ is expressed through the conditional probability $p(\mathcal{D}|\boldsymbol{w})$**.**
  - Then the uncertainty in **w** after we have observed $\mathcal{D}$ in the form of the posterior probability $p(\boldsymbol{w}|\mathcal{D})$:

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})}.$$

- $p(\mathcal{D}|\boldsymbol{w})$ is a function of parameter vector $\boldsymbol{w}$ called the *likelihood function* expressing how probable the observed data set is for different settings of the parameter vector $\boldsymbol{w}$.
- The denominator $p(\mathcal{D})$ is

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})\, d\boldsymbol{w}.$$

1.4.4. Sum Rule, Product Rule, and Bayes' Theorem

- Given the definition of likelihood, we can state Bayes' theorem in words

  $$\text{Posterior} \propto \text{likelihood} \times \text{prior}$$

  where all of these quantities are viewed as functions of $\boldsymbol{w}$.

- In the sum and product rule approach, $\boldsymbol{w}$ is considered to be a fixed parameter found by some 'estimator' and errors are obtained by considering the distribution of possible data sets $\mathcal{D}$.
- The common estimator is *maximum likelihood* in which $\boldsymbol{w}$ is set to the value that maximizes $p(\mathcal{D}|\boldsymbol{w})$. In ML, we use the negative log of the likelihood function and call it the *error function* – it is a monotonically decreasing function. Maximizing the likelihood is equivalent to minimizing the error.

- In Bayesian approach, the uncertainty in the parameters is expressed through the probability distribution over $\boldsymbol{w}$ in a single data set $\mathcal{D}$. Therefore, the inclusion of prior knowledge arises naturally.

- Example: Suppose that a fair-coin is tossed three times and lands heads each time. What is the probability of landing heads?

1.4.4. Sum Rule, Product Rule, and Bayes' Theorem

Question1: what is the probability of picking an apple?

$$p(F=a)=p(F=a|B=r)p(B=r)+p(F=a|B=b)p(B=b)$$
$$=\frac{1}{4}\left(\frac{4}{10}\right)+\frac{3}{4}\left(\frac{6}{10}\right)=\frac{11}{20}.$$

Question2: what is the probability of picking an orange from the blue box?

$$p(B=b|F=o)=1-p(r|o)=1-\frac{p(o|r)p(r)}{p(o)}$$
$$=1-\left(\frac{3}{4}\left(\frac{4}{10}\right)\left(\frac{20}{9}\right)\right)=\frac{1}{3}$$

1.4.4. Sum Rule, Product Rule, and Bayes' Theorem

- **Prior and Posterior Probability**: The probability available before we observe the identity is called the prior probability and the probability after the observation is called the posterior probability. In this example, $p(B)$ is the prior probability and $p(B|F)$ is the posterior probability.

- **Independent**: If the joint distribution of two variables factorizes into the product of the marginals, so that
$$p(X,Y) = p(X)p(Y)$$
then, X and Y are *independent*. If each box had the same fraction of apples and oranges, then
$$p(F|B) = p(F).$$

1.4.5. Expectations and Covariances

- The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the *expectation* of $f(x)$ and will be denoted as $\mathbb{E}(f)$

$$\mathbb{E}(f) = \sum_x p(x)f(x)$$

- So that the average is weighted by the relative probabilities of the different values of x. In the case of continuous variables, expectations are expressed in terms of an integration w.r.t. the corresponding probability density

$$\mathbb{E}(f) = \int p(x)f(x)dx\,.$$

- For a case of several variables,

$$\mathbb{E}_x\big(f(x,y)\big)$$

where the subscript x is the variable being averaged over and $f(x,y)$ is the function w.r.t. the distribution of x.

- The *conditional expectation* w.r.t. a conditional distribution is

$$\mathbb{E}_x(f|y) = \sum_x p(x|y)f(x)\,.$$

1.4.5. Expectations and Covariances

- The *variance* of $f(x)$ is defined by

$$var(f) = \mathbb{E}\left[\left(f(x) - \mathbb{E}\big(f(x)\big)\right)^2\right]$$

- And provides a measure of how much variability there is in $f(x)$ around its $\mathbb{E}\big(f(x)\big)$. The variance also can be written as $\mathbb{E}(f(x)^2) - \mathbb{E}\big(f(x)\big)^2$:

$$\begin{aligned}\mathbb{E}\left[\left(f(x) - \mathbb{E}\big(f(x)\big)\right)^2\right] &= \mathbb{E}\left[f(x)^2 - 2f(x)\mathbb{E}\big(f(x)\big) + \mathbb{E}\big(f(x)\big)^2\right] \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}\big(f(x)\big)\mathbb{E}\big(f(x)\big) + \mathbb{E}\left[\mathbb{E}\big(f(x)\big)^2\right] \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2.\end{aligned}$$

- For two random variables x and y, the *covariance* is the product of their deviations from their respective means,

$$cov[x, y] = \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

when x and y vary together. Otherwise, the covariance vanishes. We also can consider $cov(x) = cov[x, x]$.

1.4.5. Expectations and Covariances

- The *correlation* between two random variables is the normalized covariance between them via

$$corr(x,y) = \frac{cov[x,y]}{\sqrt{var(x)var(y)}}.$$

- The *empirical mean* vector is the arithmetic average of the observation for each variable and is defined as

$$\boldsymbol{\mu} = \overline{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n\,.$$

- The *empirical covariance* is

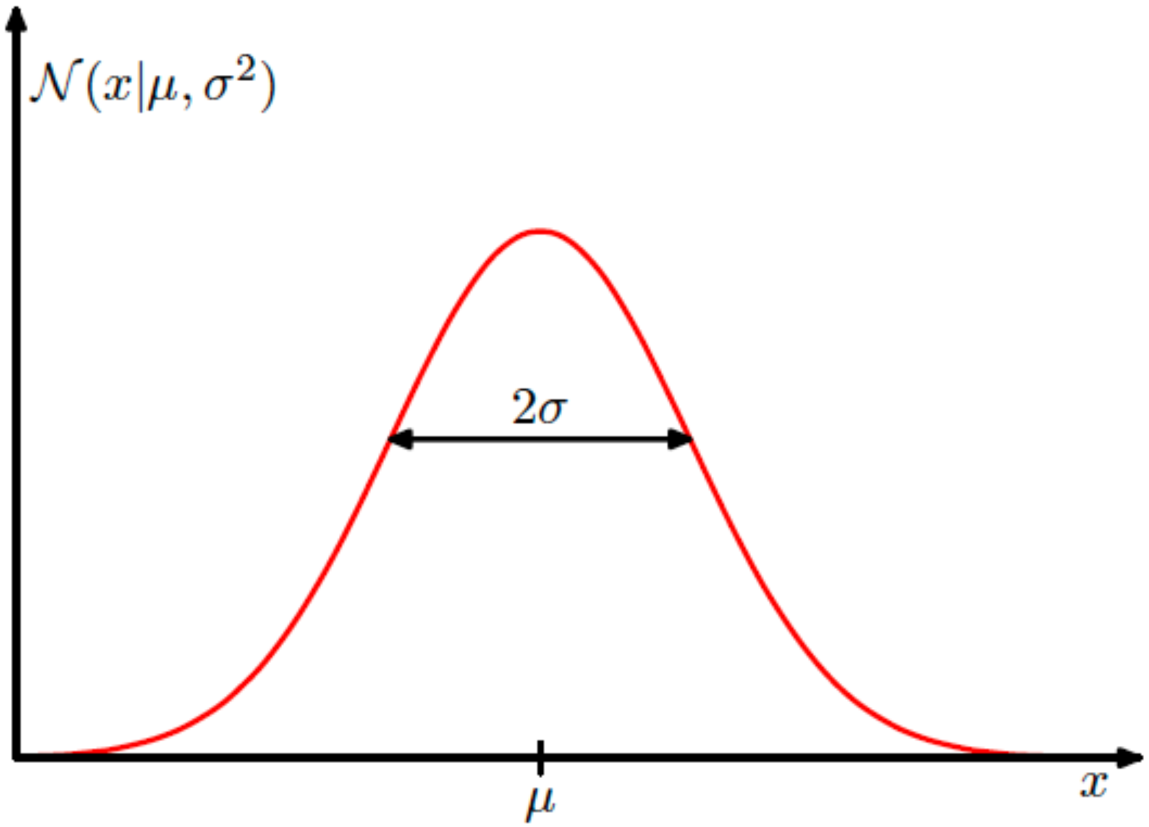
$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \overline{\boldsymbol{x}})(\boldsymbol{x}_{\boldsymbol{n}} - \overline{\boldsymbol{x}})^T\,.$$

1.4.6. The Gaussian Distribution

For the case of a single real-valued variable x, the Gaussian distribution is defined by

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

where μ is the mean and σ^2 is the variance ($\sqrt{\sigma^2}$ is the standard deviation). Sometimes, we use $\beta = 1/\sigma^2$ and it is called the precision.



1.4.6. The Gaussian Distribution

Properties:

- Condition

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

- Normalization

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\, dx = 1$$

- Expectation value

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)x\, dx = \mu$$

- Variance

$$var[x] = \sigma^2$$

- The maximum of a distribution is the mode that is coincides with the mean.

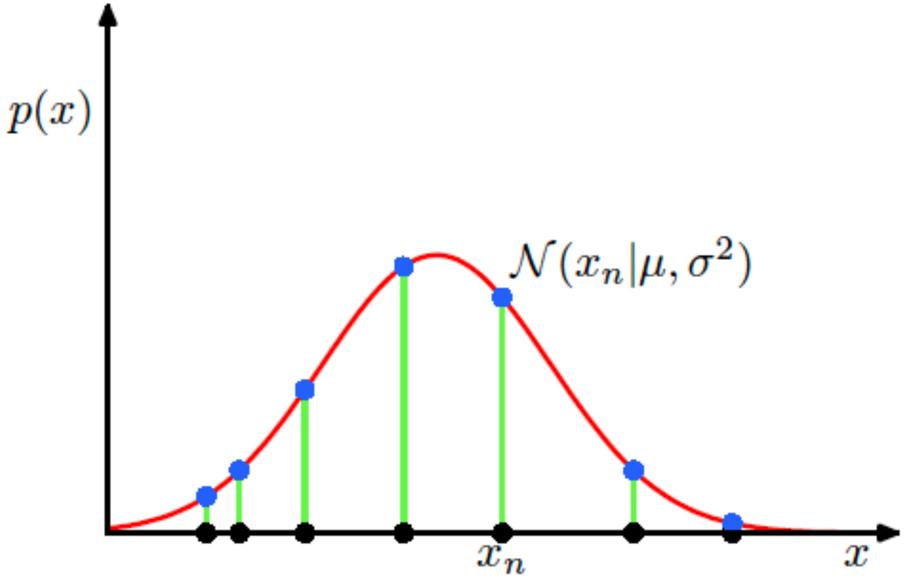
1.4.6. The Gaussian Distribution

The Gaussian distribution over a D-dimensional vector $\boldsymbol{x}$ of continuous variables is given by

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \left(\frac{1}{2\pi}\right)^{\frac{D}{2}} \left(\frac{1}{|\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

Suppose we have a data set of observations $X = (x_1, \dots, x_N)^T$ and the observations are drawn independently from a Gaussian distribution from unknown μ and σ^2. We call this situation *independent and identically distributed* (i.i.d.). The probability of the data set is then

$$p(X|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2).$$



1.4.6. The Gaussian Distribution

We can determine values for the unknown parameters by maximizing the log-likelihood function.

$$\ln p(X|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\,.$$

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$$

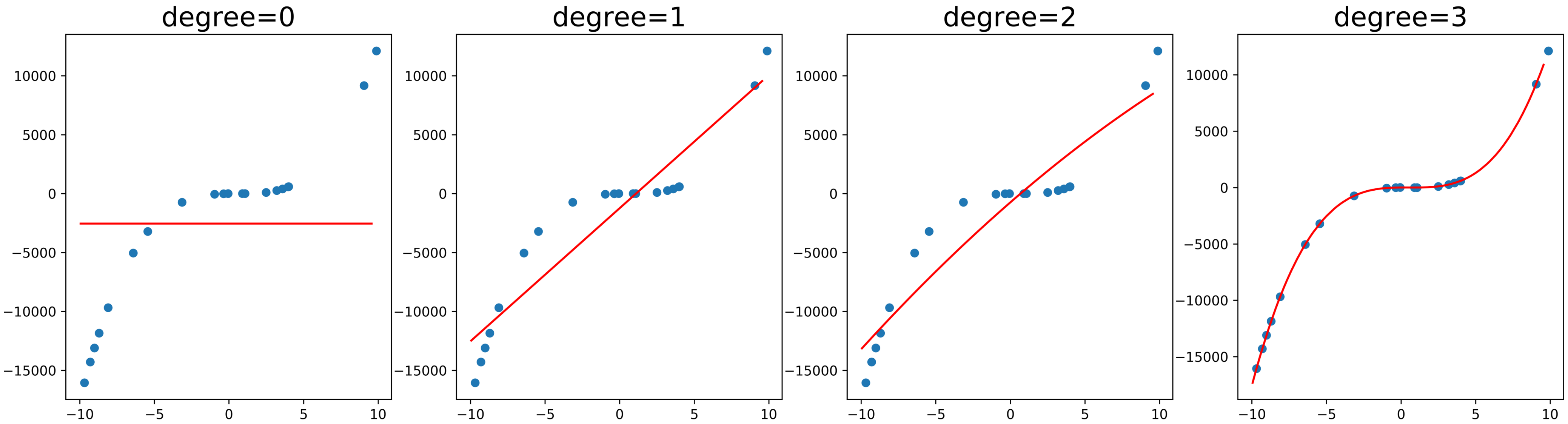
which are the *sample mean* and *sample variance*, respectively. The maximum likelihood approach systematically underestimates the variance of the distribution.

The expectation of these parameters is

$$\mathbb{E}[\mu_{ML}] = \mu$$

$$\mathbb{E}[\sigma_{ML}^2] = \frac{N-1}{N}\sigma^2.$$

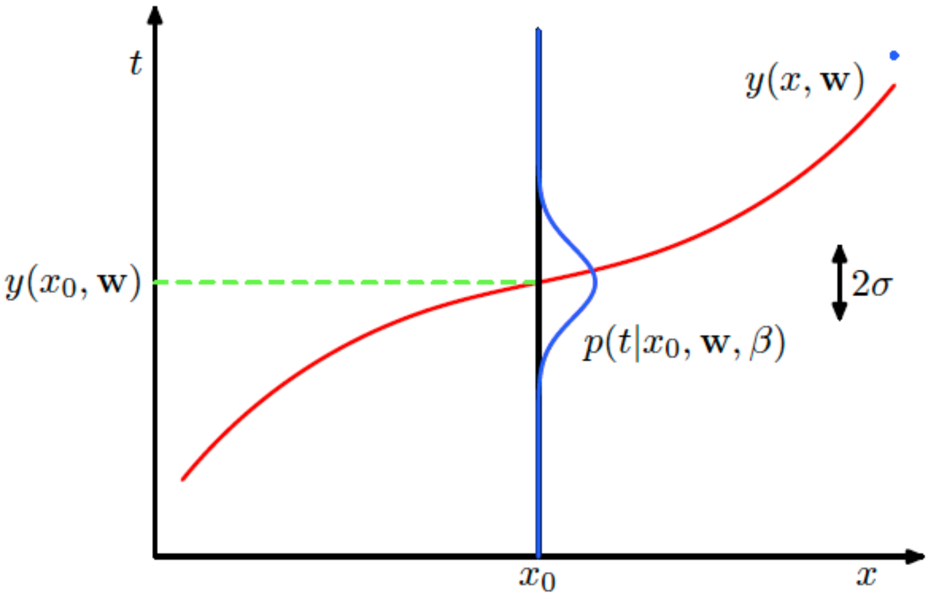
1.4.7. Curve Fitting Revisit



Suppose we predict for the target variable t given some new value of the input variable x on the basis of a set of data, $X = (x_1, \ldots, x_N)^T$ and $T = (t_1, \ldots, t_N)^T$. Assume that the given value of x, the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, \mathbf{w})$ of the polynomial curve.

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$

1.4.7. Curve Fitting Revisit



If the data are assumed to be i.i.d., then the likelihood function is

$$p(T|X, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \boldsymbol{w}), \beta^{-1})$$

and the log likelihood function is

$$\ln p(T|X, \boldsymbol{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi\,.$$

If w and β are determined, then we have a probabilistic model that are expressed in terms of the predictive distribution that gives the probability distribution over t, rather than simply a point estimate, and is obtained by substituting the maximum likelihood parameters

$$p(t|x, \boldsymbol{w}_{\boldsymbol{ML}}, \beta_{ML}) = \mathcal{N}(t\big|y(x, \boldsymbol{w}_{\boldsymbol{ML}}), \beta_{ML}^{-1}).$$

1.4.7. Curve Fitting Revisit

If we use Bayesian approach with a prior distribution over the polynomial coefficients w,

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) = \left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}} \exp\left\{-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right\}$$

where α is the precision of the distribution (called *hyperparameter* that controls the distribution of model parameter), and $M+1$ is the total number of elements in the vector w for an M^{th} order polynomial. The posterior distribution for $\boldsymbol{w}$ is

$$p(\boldsymbol{w}|X, T, \alpha, \beta) \propto p(T|X, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha).$$

We can find w by maximizing the posterior distribution called *maximum posterior* (MAP) by taking the negative log. The maximum of the posterior is equivalent to the minimum of the negative log of the posterior as shown

$$-\ln p(\boldsymbol{w}|X, T, \alpha, \beta) = -\ln p(T|X, \boldsymbol{w}, \beta) - \ln p(\boldsymbol{w}|\alpha).$$

1.4.8. Binary Variables

Consider a single binary random variable $x \in \{0,1\}$. The probability of $x = 1$ will be denoted by the parameter μ so that

$$p(x = 1|\mu) = \mu$$

where $0 \le \mu \le 1$, from which it follows that $p(x = 0|\mu) = 1 - \mu$. The probability distribution (*Bernoulli* distribution) over x can therefore be written in the form

$$Bern(x|\mu) = \mu^x (1 - \mu)^{1-x}.$$

This distribution is normalized and has mean and variance given by

$$\mathbb{E}[x] = \mu$$
$$var[x] = \mu(1 - \mu).$$

Suppose we have a data set $D = \{x_1, \dots, x_N\}$. Assume that observations are i.i.d. from $p(x|\mu)$, so that

$$p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n} (1 - \mu)^{1-x_n} \ .$$

The log likelihood is

$$\ln p(D|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\}.$$

1.4.9. Multinomial Variable

Consider the Bernoulli distribution to an K-dimensional binary variable $x_k \in \{0,1\}$ such that $\sum_k x_k = 1$. Then the distribution of **x** is given

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

where $\mu = (\mu_1, \ldots, \mu_K)^T$, and the parameters μ_k are constrained to satisfy $\mu_k \geq 0$ and $\sum_k \mu_k = 1$.

1.5. Conclusion

1.5.1. Conclusion

- A brief review of Mathematics that will play central roles in Machine Learning Algorithms
  - Linear Algebra, Analytic Geometry, and Vector Calculus will help to understand the work of algorithms.
  - Probability theory will help to understand the characteristics of algorithms
  - If discussed topics are not fully digestible, it is still okay.
    - Most of the terms will be repeatedly discussed throughout the semesters.