CS 559: Machine Learning Fundamentals & Applications

Lecture 1: Linear Algebra and Probability Theory

In Jang: ijang@stevens.edu

Spring 2021

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

- Software: Python, Jupyter notebook
- https://jupyter.org/install

Linear Algebra Review

- Scalars, Vectors, and Matrices
- Adding Matrices and Vectors
- Multiplying Matrices and Vectors
- Identity and Inverse Matrices
- Linear Dependence and Span
- Norms
- Special Matrices and Vectors
- Matrix Representation
- Linear Operators
- Trace
- Determinant
- Eigenvalues and Eigenvectors
- Subspace

Scalars, Vectors, and Matrices

- **Scalars**: a single number. (x)
  - Let $s \in \mathbb{R}$ be the slope of the line: real-valued scalar
  - Let $n \in \mathbb{N}$ be the number of units: natural number scalar
- **Vectors**: an array of numbers that are arranged in order. ($\boldsymbol{x}$)
  - Elements of vectors: typically written as x_i where x is scalar and the subscript i is the order.
  - If each element is in $\mathbb{R}$ and the vector has n elements, then the vector lies in $\mathbb{R}^n$
  - $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ or $\boldsymbol{x} = [x_1 \quad \cdots \quad x_n]$
- **Matrix**: a 2-D array of numbers. ($\boldsymbol{A}$)
  - $\boldsymbol{A} \in \mathbb{R}^{m \times n}$: A real-valued matrix $\boldsymbol{A}$ has a height of m and a width of n
  - $\boldsymbol{A} = \begin{bmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{bmatrix}$
  - Vectors can be considered as matrices that contain only one column.

Scalars, Vectors, and Matrices

- **Transpose**: The mirror image of the matrix across a diagonal line, called the main diagonal, running down and to the right, staring from its upper left corner.
  - $\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix}^T = \begin{bmatrix} a_{1,1} & a_{2,1} & a_{3,1} \\ a_{1,2} & a_{2,2} & a_{3,2} \\ a_{1,3} & a_{2,3} & a_{3,3} \end{bmatrix}$
  - $\boldsymbol{A}_{i,j}^T = \boldsymbol{A}_{j,i}$
  - If $\boldsymbol{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$, then its transpose $\boldsymbol{a}^T$ is $\boldsymbol{a}^T = [a_1 \quad \cdots \quad a_n]$

$$A\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \qquad A^T\begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

$$A\begin{bmatrix} 1 & 4 & 3 \\ 8 & 2 & 6 \\ 7 & 8 & 3 \\ 4 & 9 & 6 \\ 7 & 8 & 1 \end{bmatrix} \qquad A^T\begin{bmatrix} 1 & 8 & 7 & 4 & 7 \\ 4 & 2 & 8 & 9 & 8 \\ 3 & 6 & 3 & 6 & 1 \end{bmatrix}$$

Adding Matrices and Vectors

- Addition between matrices?
  - As long as the shapes are the same,
    - $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$ where $C_{i,j} = A_{i,j} + B_{i,j}$
- A scalar to Matrix?
  - Perform the operations to each element
    - $\boldsymbol{D} = a \cdot \boldsymbol{B} + c$ where $D_{i,j} = aB_{i,j} + c$
- A vector to Matrix?
  - Yields another matrix
    - $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{b}$ where $C_{i,j} = A_{i,j} + b_j$
    - Copying of **b** to many locations is called **broadcasting.**

The vector b is added to each row of the matrix.

Multiplying Matrices and Vectors

- Multiplication (**Matrix product**) between matrices?
  - **A** must have the same number of columns as **B** has rows
    - Let **A** be $m \times n$ and **B** be $n \times p$, then $\boldsymbol{C} = \boldsymbol{AB}$ is a matrix with shape of $m \times p$
    - $\boldsymbol{C} = \boldsymbol{AB}$ where $C_{i,j} = \sum_k A_{i,k} B_{k,j}$
    - The product of individual elements is called **element-wise product**, or **Hadamard product** and is denoted as $\boldsymbol{A} \odot \boldsymbol{B}$.
- The dot product between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ must have the same dimension
  - Also can be considered as matrix product with the same dimensionality
  - $\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{y}$

$$\begin{bmatrix} -4 & 4 & 6 \\ 2 & 3 & -1 \end{bmatrix} \begin{bmatrix} 0 & -4 & -2 & 5 \\ 1 & 6 & -1 & 7 \\ 8 & 2 & 4 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} -4\times0 + 4\times1 + 6\times8 & -4\times-4 + 4\times6 + 6\times2 & -4\times-2 + 4\times-1 + 6\times4 & -4\times5 + 4\times7 + 6\times3 \\ 2\times0 + 3\times1 + -1\times8 & 2\times-4 + 3\times6 + -1\times2 & 2\times-2 + 3\times-1 + -1\times4 & 2\times5 + 3\times7 + -1\times3 \end{bmatrix}$$

$$= \begin{bmatrix} 52 & 52 & 28 & 26 \\ -5 & 8 & -11 & 28 \end{bmatrix}$$

Multiplying Matrices and Vectors

- **Distributive property:**
  - $\boldsymbol{A(B+C) = AB + AC}$
- **Associative property:**
  - $\boldsymbol{A(BC) = (AB)C}$
- **But not commutative!**
  - $\boldsymbol{AB \neq BA}$
  - but in dot product between two vectors, we have $\boldsymbol{x^T y = y^T x}$.

Identity and Inverse Matrices

Matrix inversion: Enables to analytically solve $\boldsymbol{Ax} = \boldsymbol{b}$ for many values of $\boldsymbol{A}$.

- An identity matrix: a matrix that does not change any vector when we multiply that vector by that matrix. It is denoted as $\boldsymbol{I_n}$ that is $\in \mathbb{R}^{n \times n}$ and for $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{I_n x} = \boldsymbol{x}$.
  - All the entries along the main diagonal are 1 and other entries are zero.
- The matrix inverse of $\boldsymbol{A}$ is denoted as $\boldsymbol{A^{-1}}$ and it is defined as the matrix such that

$$\boldsymbol{A^{-1}A} = \boldsymbol{I_n}$$

If we have $\boldsymbol{Ax} = \boldsymbol{b}$, $\boldsymbol{x} = \boldsymbol{A^{-1}b}$.

- $\boldsymbol{A^{-1}}$ is primarily useful as a theoretical tool but *should not be used* in practice for most software applications.
- $\boldsymbol{A^{-1}}$ can be represented with only limited precision on a digital computer and algorithms that make use of the value of $\mathbf{b}$ can be usually obtain more accurate estimates of $\boldsymbol{x}$.

The eight axioms

Associativity of addition	$\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$
Commutativity of addition	$\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
Identity element of addition	There exists an element $0 \in \mathbf{V}$, called the zero vector, such that $\mathbf{v} + 0 = \mathbf{v}$ for all $\mathbf{v} \in V$.
Inverse elements of addition	For every $v \in V$, there exists an element $-\mathbf{v} \in V$, called the additive ***inverse*** of $\mathbf{v}$, such that $v + (-v) = 0$.
Compatibility of scalar multiplication with field multiplication	$a(b\mathbf{v}) = (ab)\mathbf{v}$
Identity element of scalar multiplication	$1v = v$, where 1 denotes the multiplicative identity in F.
Distributivity of scalar multiplication with respect to vector addition	$a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$
Distributivity of scalar multiplication with respect to field addition	$(a + b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}$

Linear Dependence and Span

Recall $\boldsymbol{Ax} = \boldsymbol{b} \rightarrow \boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$

- One solution for every value of b but also can have no solutions or infinitely many solutions
- It is not possible to have more than one but less than infinitely many solutions for particular b.
- If both x and y are solutions, then

$$\boldsymbol{z} = \alpha\boldsymbol{x} + (1 - \alpha)\boldsymbol{y}$$

Linear Dependence and Span

The columns of **A** specify directions from the origin

- Determines *how many ways* there are of reaching $\boldsymbol{b}$.
- **Linear combination**

$$\boldsymbol{Ax} = \sum_i x_i A_{:,i}$$

- **Span** – set of all points obtained by linear combination of the original vectors.
- Determining whether $\boldsymbol{Ax} = \boldsymbol{b}$ has a solution is testing whether $\boldsymbol{b}$ is in the *span of the columns* of $\boldsymbol{A}$.
  - The column space of $\boldsymbol{A}$ be all of $\mathbb{R}^m$ for the system to have solution for all values of $\boldsymbol{b} \in \mathbb{R}^m$!
  - $\boldsymbol{A}$ must have at least m columns!

Linear Dependence and Span

Consider A have columns n and b has m elements where $n \geq m$,

- If A is a 3×2 matrix and x is 2×1, the target b is a 3×1.
- The target $\boldsymbol{b}$ is in 3-D but x is only 2-D.
- In this case, the solution is only when $\boldsymbol{b}$ lies on that plane.
- What happens if columns are identical?
  - Wil fail to encompass all of $\mathbb{R}^m$
  - **Linear independent** – if no vector in the set is a linear combination of the other vectors.
  - For the column space of the matrix to encompass all of $\mathbb{R}^m$, the matrix must contain at least one set of m linearly intendent columns.
  - This requirement is to have exactly m linear independent columns, not at least m.
  - No set of m-D vectors can have more than m mutually independent columns, but a matrix with more than m columns may have more than one such set.

Linear Dependence and Span

- $\boldsymbol{Ax = b \rightarrow x = A^{-1}b}$

For A to have an inverse:

- Linear independency
- Must have at most one solution for each value of b → at most m columns
  - Otherwise, there is more than one way of parameterizing each solution.

- In other words… A must be **square** matrix by being $m \times m$ and all columns be **linearly independent**.
- A square matrix with linearly dependent columns is known as **singular.**

- For non-square or a square but singular **A?**
  - **We can solve using other than the method of matrix inversion.**

Norms

Norm – In ML, we measure the size of vectors by mapping vectors to non-negative values.

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

for $p \in \mathbb{R}, p \geq 1$.

Also the norm of vector x measures the distance from the origin to the point *x*.

Matrix Representation

- Let $\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}$ be a basis
- Every $\mathbf{v}$ can be uniquely represented as:

$$\mathbf{v} = \alpha_1\mathbf{v}_1 + \cdots + \alpha_k\mathbf{v}_k$$

- Denote $\mathbf{v}$ by the column-vector: $\mathbf{v} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}$
- Denote the basis vectors as: $\begin{pmatrix} \mathbf{1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{1} \\ \vdots \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{1} \end{pmatrix}$

Linear Operators

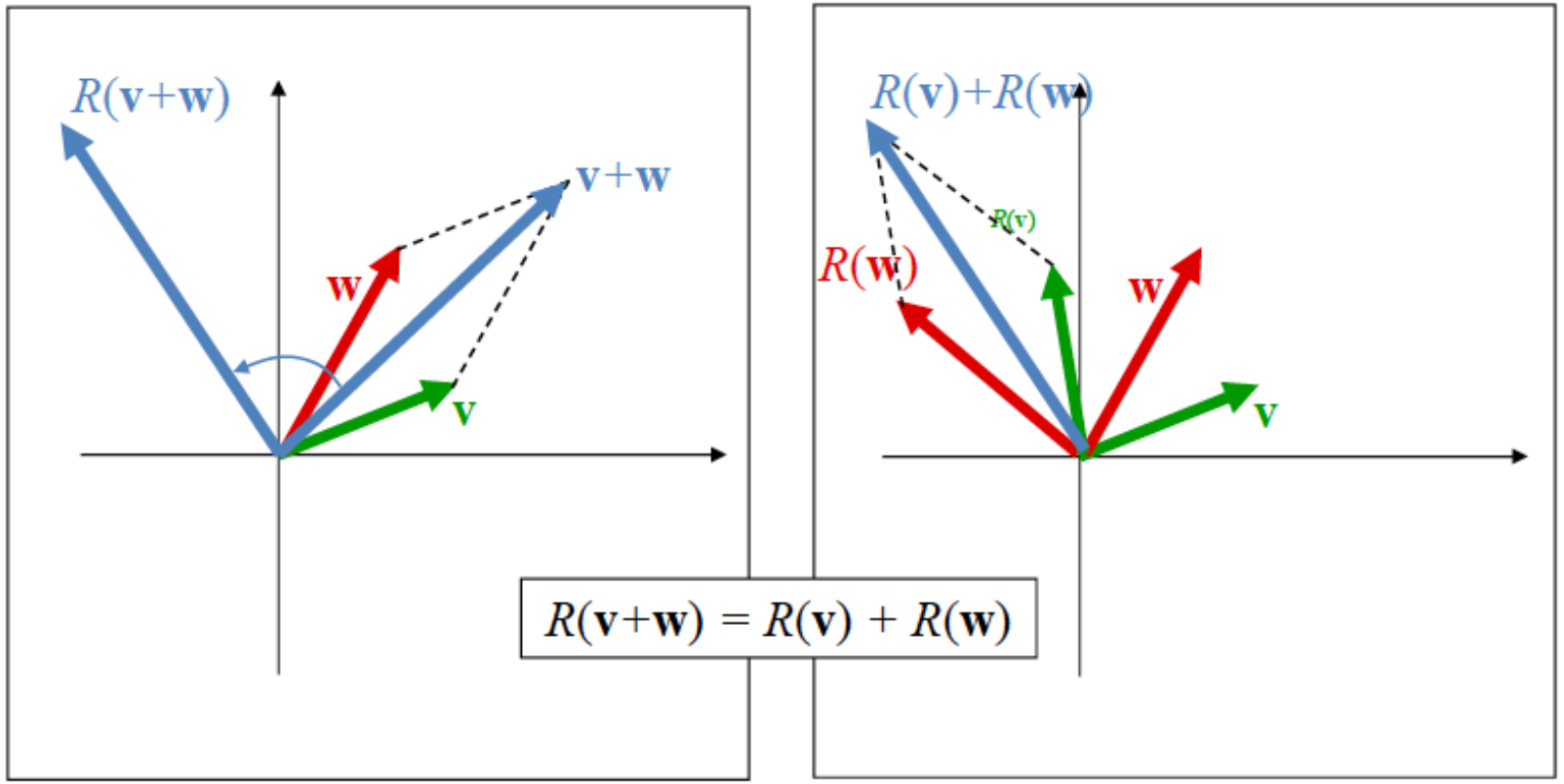
- $A : \mathrm{V} \rightarrow \mathrm{W}$ is called linear operator if:
  - $A(\mathbf{v} + \mathbf{w}) = A(\mathbf{v}) + A(\mathbf{w})$
  - $A(\alpha\mathbf{v}) = \alpha A(\mathbf{v})$

- In particular, $A(\mathbf{0}) = 0$

- Are the following operators linear?
  - Scaling
  - Rotation
  - Translation

Linear operators - illustration

- Rotation is a linear operator:



Matrix properties

- Matrix $A_{(n\times n)}$ is **non-singular** if $\exists B,\ AB = BA = I$
- $B = A^{-1}$ is called the **inverse** of A
- A is non-singular $\Leftrightarrow$ $det\ A \neq 0$

- If A is non-singular then the equation
  - $A\mathbf{x}=\mathbf{b}$ has one unique solution for each $\mathbf{b}$
  - the rows of A are linearly independent (and so are the columns)

Orthogonal matrices

- Matrix $A_{(n\times n)}$ is orthogonal if $A^{-1} = A^T$
- Follows: $AA^T = A^TA = I$
- The rows of A are orthonormal vectors!

  Proof:

$$I = A^TA = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \\ \mathbf{v}_n \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & & \mathbf{v}_n \end{pmatrix} = \begin{pmatrix} \mathbf{v}_i^T\mathbf{v}_j \end{pmatrix} = \begin{pmatrix} \delta_{ij} \end{pmatrix}$$

$$\Rightarrow \langle \mathbf{v}_i, \mathbf{v}_i \rangle = 1 \quad \Rightarrow \quad \|\mathbf{v}_i\| = 1; \quad \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$$

Trace

- The trace of a square matrix denoted by tr(A) is sum of the diagonal elements

$$\mathrm{tr(A)} = \sum_{i=1}^{N} A_{ii}$$

- $tr\left(\begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}\right) = A_{11} + A_{22} = 1 + 1 = 2$

Determinant

- For a square matrix A, the determinant is denoted by $|A|$ or $\det(A)$

$$
\begin{aligned}
\left|\left[a_{11}\right]\right| &= a_{11} \\
\left|\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right| &= a_{11}a_{22} - a_{12}a_{21} \\
\left|\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right| &= \begin{aligned}[t] & a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ & -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{aligned}
\end{aligned}
$$

Determinant

- $|A| = |A^{\mathrm{T}}|$
- $|AB| = |A|\,|B|$
- $|A| = 0$, if and only if A is singular
  - Else, $|A^{-1}| = 1/|A|$

Eigenvalues and Eigenvectors

- For an n×n **square** matrix A, e is an eigenvector with eigenvalue λ if

$$Ae = \lambda e$$

- Or

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

- If $(A - \lambda \mathbf{I})$ is invertible, the only solution is e=0 (trivial)

Eigenvalues and Eigenvectors

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

- For non-trivial solutions:

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

- Above equation is called the “characteristic polynomial”
- Solutions are not unique
  - If e is an eigenvector αe is also an eigenvector

Simple Example

- For a 2×2 matrix

$$\det[\mathbf{A}-\lambda\mathbf{I}]=\begin{vmatrix}a_{11}-\lambda & a_{12}\\ a_{21} & a_{22}-\lambda\end{vmatrix}=(a_{11}-\lambda)(a_{22}-\lambda)-a_{12}a_{21}=0$$

$$0=a_{11}a_{22}-a_{12}a_{21}-\lambda(a_{11}+a_{22})+\lambda^2$$

Simple Example

$$0 = a_{11}a_{22} - a_{12}a_{21} - (a_{11} + a_{22})\lambda + \lambda^2$$

$$0 = 1 \cdot 4 - 2 \cdot 2 - (1 + 4)\lambda + \lambda^2$$

$$(1 + 4)\lambda = \lambda^2$$

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

- The solutions are λ=0 and λ=5
- The eigenvector for the first eigenvalue, λ=0 is:

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

$$\left[\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\right] \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1x + 2y \\ 2x + 4y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- One solution for both equations is x=2, y=-1

Simple Example

- The second eigenvalue is λ=5

$$\left[\begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}-\begin{bmatrix}5 & 0\\ 0 & 5\end{bmatrix}\right]\cdot\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}-4 & 2\\ 2 & -1\end{bmatrix}\cdot\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}-4x+2y\\ 2x-1y\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}$$

$$-4x+2y=0,\ \text{and}\ 2x-y=0,\ \text{so},\ x=1, y=2$$

Properties

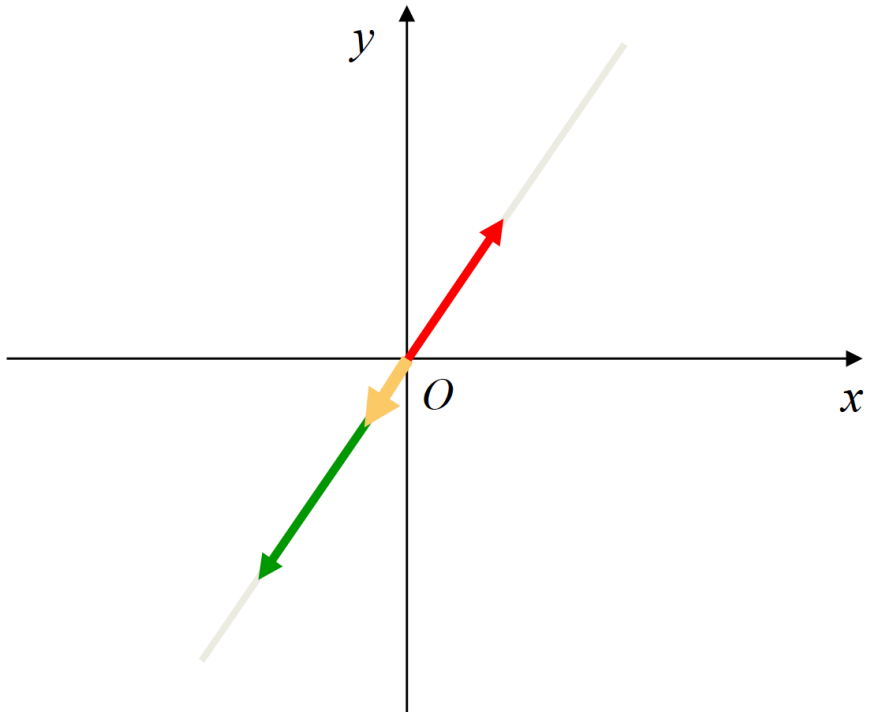
- The product of the eigenvalues = |A|
- The sum of the eigenvalues = trace(A)
- The eigenvectors are pairwise orthogonal

Subspace

- Let F be a field, V be a vector space over F, and let W be a subset of V. Then W is a **subspace** if:
  - The zero vector, $\mathbf{0}$, is in W.
  - If $\mathbf{u}$ and $\mathbf{v}$ are elements of W, then the sum $\mathbf{u} + \mathbf{v}$ is an element of W.
  - If $\mathbf{u}$ is an element of W and c is a scalar from K, then the scalar product $c\mathbf{u}$ is an element of W.

Subspace Example

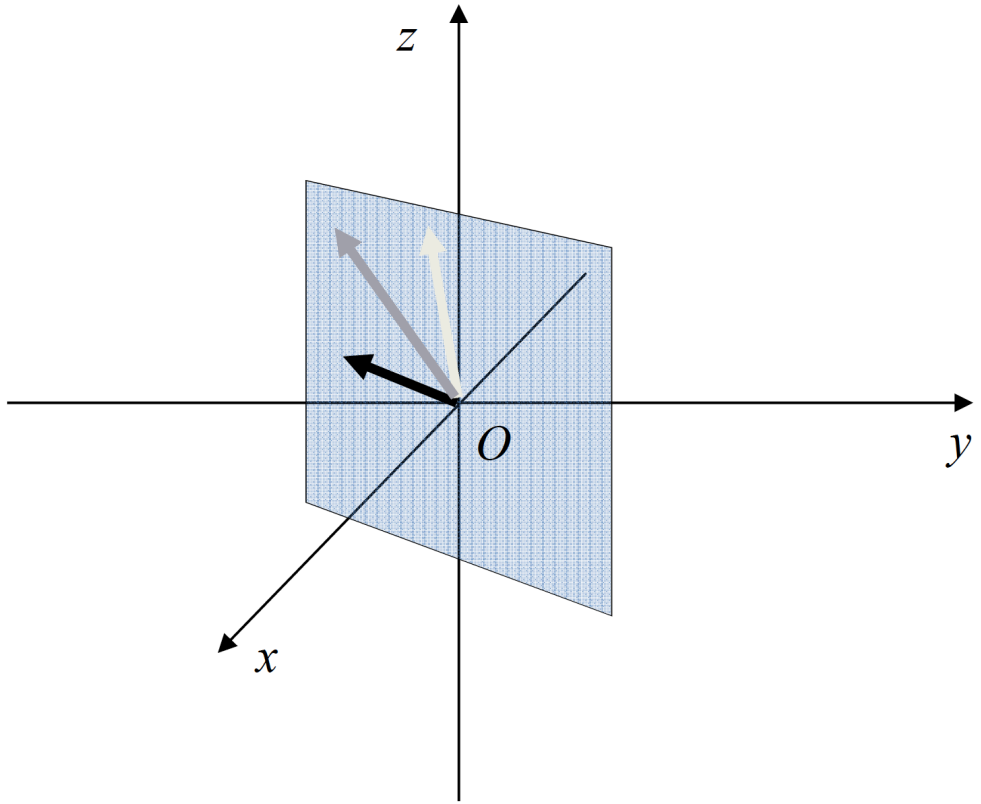
- Let l be a 2D line though the origin
- $L = \{\mathbf{p} - O \mid \mathbf{p} \in l\}$ is a linear subspace of R^2



O. Sorkine, 2006

Subspace Example

- Let π be a plane through the origin in 3D
- $V = \{\mathbf{p} - O \mid \mathbf{p} \in \pi\}$ is a linear subspace of R^3



Linear Independence

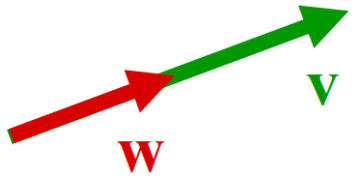
- The vectors $\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}$ are a linearly independent set if:

$$\alpha_1\mathbf{v}_1 + \cdots + \alpha_k\mathbf{v}_k = \mathbf{0} \Leftrightarrow \alpha_i = 0\ \forall i$$

- It means that none of the vectors can be obtained as a linear combination of the others.

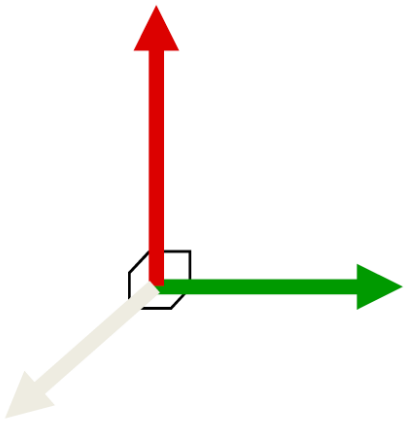
Linear independence - example

- Parallel vectors are always dependent:



$$\mathbf{v} = 2.4\,\mathbf{w} \quad \Rightarrow \quad \mathbf{v} + (-2.4)\mathbf{w} = 0$$

- Orthogonal vectors are always linearly independent:



Basis of Vector Space

- $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ are **linear independent**
- $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ **span** the whole vector space $\boldsymbol{V}$

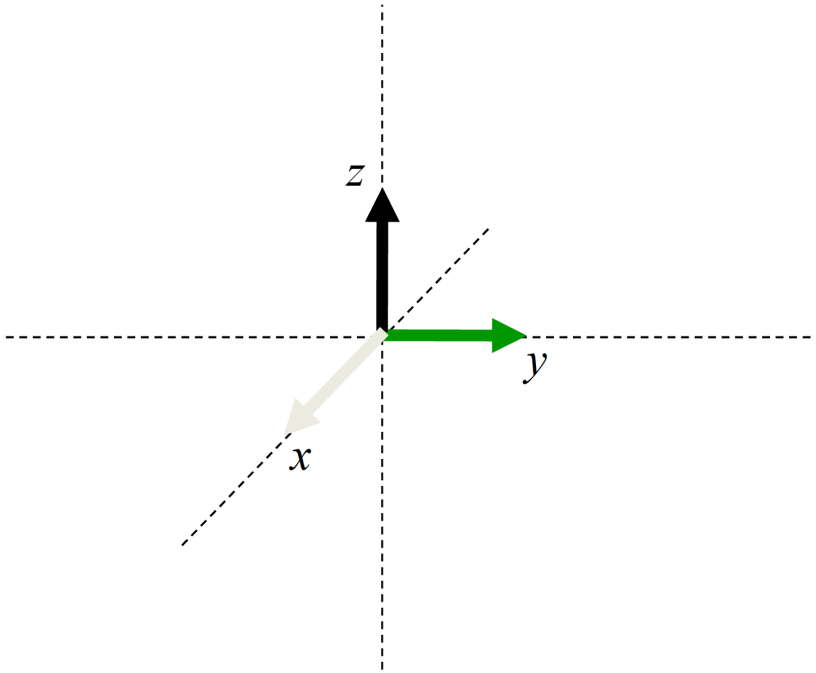
$$\mathbf{V} = \{\alpha_1 \mathbf{v}_1 + \cdots + \alpha_k \mathbf{v}_k | \alpha_1 \in R\}$$

- Any vector in $\mathbf{V}$ is a **unique** linear combination of the basis

- The number of basis vectors is called the **dimension** of $\mathbf{V}$

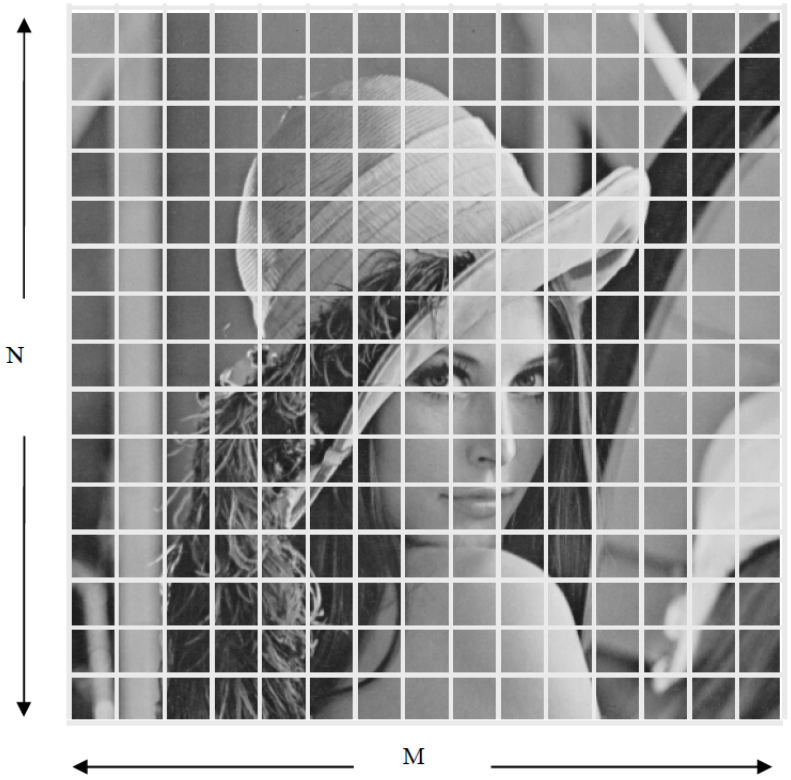
Basis Example

- The standard basis of $\mathbb{R}^3$



Basis – Another Example

- Grayscale $N \times M$ images:
  - Each pixel has value between 0 (black) and 1 (white)
  - The image can be interpreted as a vector $\in \mathbb{R}^2$



Probability Theory

- Axioms of Probability
- Discrete Random Variables
- Continuous Random Variables
- The Univariate Gaussian
- Maximum-likelihood Estimator

The Axioms of Probability

- P(A) = # of event/# of possible event
- 0 <= P(A) <= 1
- P(True) = 1 & P(False) = 0
- P(A or B) = P(A) + P(B) – P(A and B)

Discrete Random Variables

- Sample space Ω: Possible "states" x of the random variable X (Outcomes of the experiment, output of the system, measurement)
- Either have a finite or countable number of states.
- Events: Possible combination of states ('subsets of Ω')

Probability Mass Functions (PMF)

A function which tells us how likely each possible outcome is:

$$P(X = x) = P_x(x) = P(x)$$

$$\sum_{x \in \Omega} P(x) = 1$$

$$P(A) = P(x \in A) = \sum_{x \in A} P(X = x)$$

Expectation and Variance

- Expectation (or mean):

$$E(x) = \sum_x P(X = x)x$$

- Expectation of a function:

$$E\big(f(x)\big) = \sum_x P(X = x)f(x)$$

- Moments = expectation of power of X:

$$M_k = E(X^k)$$

Expectation and Variance

- Variance: Average (squared) fluctuation from the mean

$$Var(X) = E\left(\left(X - E(X)\right)^2\right) = E(X^2) - E(X)^2$$
$$= M_2 - M_1^2$$

- Standard deviation: square root of variance $\sqrt{Var(x)} = \sqrt{M_2 - M_1^2}$

Bivariate Distributions

- Joint Distributions: $P(X = x, Y = y)$, a list of all probabilities of all possible pairs of observations
- Marginal Distribution:

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

- Conditional Distribution: $P(X = x|Y = y) = P(X = x, Y = y)$
  - $X|Y$ has distribution $P(X|Y)$, where $P(X|Y)$ specifies a "lookup-table" of all possible $P(X = x|Y = y)$

Expectation and Covariance of Bivariate Distributions

- Conditional distributions are just distributions which have a (conditional) mean or variance
- Covariance is the expected value of the product of fluctuations:

$$Cov(X,Y) = E\left(\big(X - E(X)\big)\big(Y - E(Y)\big)\right) = E(XY) - E(X)E(Y)$$

$$Var(X) = Cov(X,X)$$

- One common way to construct bivariate random variables is to have a random variable whose parameter is another random variable.

- Two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$
- $P(A)P(B) = P(A \cap B)$
- Two random variables are independent if the joint PMF is the product of the marginals:
  - $P(X = x, Y = y) = P(X = x)P(Y = y)$
- If X and Y are independent, we write $X \perp Y$. Knowing the value of X does not tell us anything about Y.
- If X and Y are independent, $Cov(X,Y)=0$.
- Mutual information is a measure of how “non-independent" two random variables are.

Multivariate Distributions

- $\boldsymbol{X}$, $\boldsymbol{x}$ are vector.
- Mean:

$$E(\boldsymbol{X}) = \sum_{x} \boldsymbol{x} P(\boldsymbol{x})$$

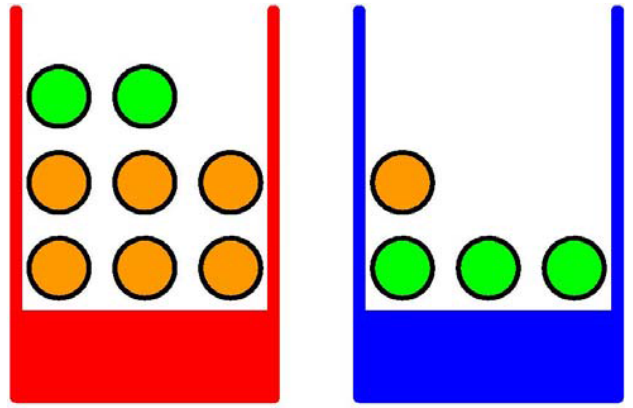
- Covariance Matrix:

$$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$$
$$Cov(\boldsymbol{X}) = E(\boldsymbol{X}\boldsymbol{X}^T) - E(\boldsymbol{X})E(\boldsymbol{X})^{\boldsymbol{T}}$$

- Conditional and marginal distributions: Can define and calculate any (multi- or single dimensional) marginals or conditional distributions we need: $P(X_1), P(X_1, X_2), P(X_1, X_2, X_3 | X_4)$, etc…

Example

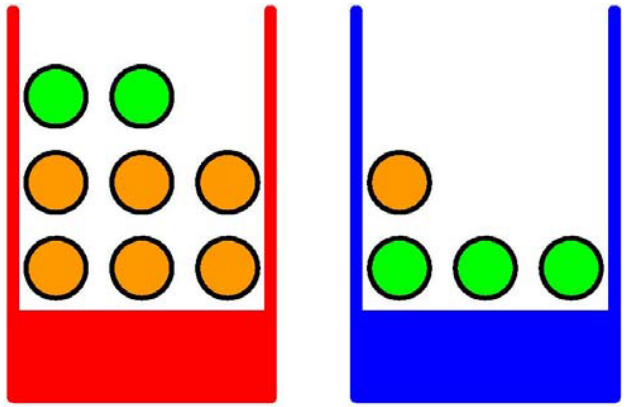
- Let us look at the following example:
  - We have two boxes, one red and one blue
  - Red box: 2 apples and 6 oranges
  - Blue box: 3 apples and 1 orange
  - Pick red box 40% of the time and blue box 60% of the time, then pick one item of fruit



C.M. Bishop, “Pattern Recognition and Machine Learning”, 2006

Example

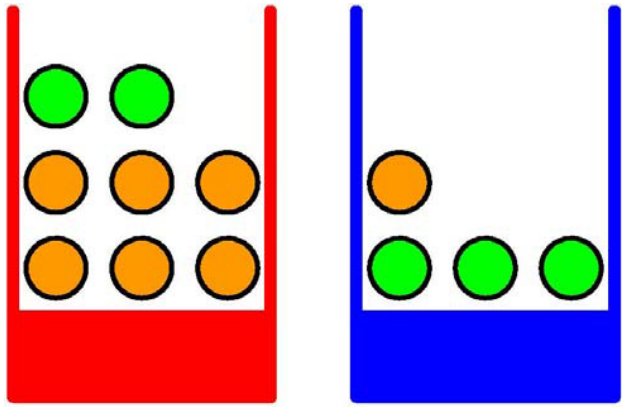
- Define:
  - B random variable for box picked
    - B = {blue(b), red(r)}
  - F identity of fruit
    - F = {apple(a), orange(o)}
  - P(B=r)=0.4 and P(B=b)=0.6
  - Events are mutually exclusive and include all possible outcomes
  - Their probabilities must sum to 1



Example

- P(B=r) = 0.4, P(B=b) = 0.6
- P(B=r) + P(B=b) = 1.0

- Conditional Probabilities
  - P(F=a|B=r) = 2/8 = 0.25
  - P(F=o|B=r) = 6/8 = 0.75
  - P(F=a|B=b) = 3/4 = 0.75
  - P(F=o|B=b) = 1/4 = 0.25



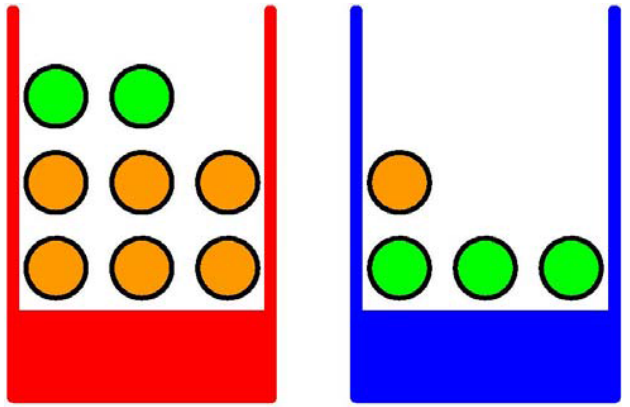
Example

- Note: P(F=a|B=r)+P(F=o|B=r) = 1

- P(F=a) = P(F=a|B=r)P(B=r)+P(F=a|B=b)P(B=b)
  = 1/4 * 4/10 + 3/4 * 6/10 = 11/20

- P(F=o) = 1-11/20=9/20=0.45

P(F=a|B=r) = 2/8 = 0.25
P(F=o|B=r) = 6/8 = 0.75
P(F=a|B=b) = 3/4 = 0.75
P(F=o|B=b) = 1/4 = 0.25



Prior vs. Posterior

- Prior Probability - If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability P(B).
- Posterior Probability - Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $P(B|F)$, which we shall call the posterior probability because it is the probability obtained after we have observed F.

Conditional Probability

- Conditional probability: Recalculated probability of event A after someone tells you that event B happened.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A|B)P(B)$$

- Bayes Rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Bayesian Probability

- Bayesian view: probabilities provide a quantification of uncertainty. Before observing the data, the assumptions about w are captured in the form of a prior probability distribution $P(\boldsymbol{w})$. The effect of the observed data $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ is expressed by $P(D|\boldsymbol{w})$.
- Bayes' theorem:

$$P(\boldsymbol{w}|D) = \frac{P(D|\boldsymbol{w})P(\boldsymbol{w})}{P(D)}$$

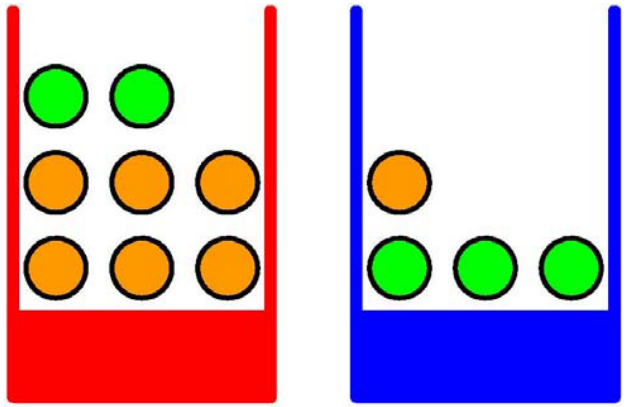
- Bayes' theorem in words: **posterior $\propto$ likelihood $\times$ prior**

Bayes Rule on the Fruit Example

- Assume we have now picked an orange
- Ask: which is the probability that it was from the red box?

$$P(B=r|F=o) = \frac{P(F=o|B=r)P(B=r)}{P(F=o)} = \frac{0.75 \times 0.4}{0.45} = \frac{2}{3}$$

P(B=r) = 0.4, P(B=b) = 0.6
P(F=o) = 0.45,
P(F=o|B=r) = 6/8 = 0.75



Continuous Random Variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_x(x)$ is a probability density function for X, then

$$P(a < X < b) = \int_a^b p(x)dx$$

$$P(a < X < a + dx) \approx p(a) \cdot dx$$

Continuous Random Variables

- The cumulative distribution function is $F_x(x) = P(X < x)$. We have that $p_x(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)ds$.
- More generally, If A is an event, then

$$P(A) = P(X \in A) = \int_{x \in A} p(x)dx$$

$$P(\Omega) = P(X \in \Omega) = \int_{x \in \Omega} p(x)dx = 1$$

Mean, Variance, and Conditionals

- Mean: $E(x) = \int_x x \cdot p(x)dx$
- Variance: $Var(X) = E(X^2) - E(X)^2$
- If X has pdf *p(x)*, then $X|(X \in A)$ has pdf

$$p_{(x|A)}(x) = \frac{p(x)}{P(A)} = \frac{p(x)}{\int_{x \in A} p(x)dx}$$

- Only makes sense if $P(A) > 0$!

Bivariate Continuous Distributions

- $p_{x,y}(x, y)$, joints probability density function of X and Y
- $\int_x \int_y p(x, y) dx dy = 1$
- Marginal distribution: $p_X(x) = \int_{-\infty}^{\infty} p(x, y) dy$
- Conditional distribution: $p(x|y) = \frac{p(x,y)}{p(y)}$
- Note: $P(Y = y) = 0$!
- Independence: X and Y are independent if $p_{(x,y)}(x, y) = p_x(x) p_y(y)$

The Univariate Gaussian

- Probability density function

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

- Easy to validate:

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

- Expectation

$$E(x) = \int_{-\infty}^{\infty} p(x|\mu,\sigma^2)xdx = \mu$$

- Variance

$$Var(x) = E(x^2) - E(x)^2 = \sigma^2$$

Products of Gaussian pdfs

- Suppose $p_1(x) = p\left(x, \mu_1, \frac{1}{\beta_1}\right)$ and $p_2(x) = p\left(x, \mu_2, \frac{1}{\beta_2}\right)$, then

$$p_1(x)p_2(x) \propto p\left(x, \mu, \frac{1}{\beta}\right)$$

$$\beta = \beta_1 + \beta_2$$

$$\mu = \frac{1}{\beta}(\beta_1\mu_1 + \beta_2\mu_2)$$

- In general,

$$p_1(x)p_2(x) \dots p_n(x) \propto p\left(x, \mu, \frac{1}{\beta}\right)$$

$$\beta = \sum_n \beta_n$$

$$\mu = \frac{1}{\beta}\sum_n \mu_n\beta_n$$

- This is also true for multivariate Gaussians!

Maximum Likelihood Estimator

- Assuming data points are independent and identically distributed (i.i.d.), the probability of the data set given μ and σ^2 (the likelihood function):

$$P(\boldsymbol{x}|\mu,\sigma^2) = \prod_{n=1}^{N} p(x_n|\mu,\sigma^2)$$

- log-likelihood:

$$\log P(\boldsymbol{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log(2\pi)$$

- Maximizing log-likelihood with respect to μ and σ^2:

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