

Basic Mathematics

A Machine Learning Perspective

S.V.N. “Vishy” Vishwanathan
vishy@axiom.anu.edu.au

National ICT of Australia
and
Australian National University

Thanks to Alex Smola for initial version of slides

- Functional Analysis
- Linear Algebra
- Matrix Theory
- Probability

Metric Space:

A pair (\mathcal{X}, d) , where \mathcal{X} is a set and $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_0^+$ is a metric space if $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$

- $d(\mathbf{x}, \mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$
- $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (**Symmetry**)
- $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ (**Triangle inequality**)

Examples:

Euclidean space

For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ we define $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$

ℓ^p -space

Space of sequences with $d(\mathbf{x}, \mathbf{y}) = (\sum_{i=1}^{\infty} |x_i - y_i|^p)^{\frac{1}{p}}$

Hilbert space

Space of sequences with $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{\infty} (x_i - y_i)^2}$

Ball:

Given $\mathbf{x}_0 \in \mathcal{X}$ and $r > 0$ we define

● $B(\mathbf{x}_0, r) = \{\mathbf{x} \in \mathcal{X} \mid d(\mathbf{x}, \mathbf{x}_0) < r\}$ (Open ball)

● $\bar{B}(\mathbf{x}_0, r) = \{\mathbf{x} \in \mathcal{X} \mid d(\mathbf{x}, \mathbf{x}_0) \leq r\}$ (Closed ball)

Open set:

A subset M of a metric space \mathcal{X} is open if it contains an open ball about each of its points

Closed set:

If the complement of M is open it is called a closed set

Examples:

● The set $(a, b) \subset \mathbb{R}$ is an open set

● The set $[a, b] \subset \mathbb{R}$ is a closed set

● The set $(a, b] \subset \mathbb{R}$ is neither open nor closed

Convergence:

A sequence $\{x_i\} \in \mathcal{X}$ is said to converge if for any ϵ there exists a x and a n_0 such that for all $n \geq n_0$ we have $d(x_n, x) \leq \epsilon$

Cauchy Series:

A sequence $\{x_i\} \in \mathcal{X}$ is a Cauchy if for any ϵ there exists a n_0 such that for all $m, n \geq n_0$ we have $d(\mathbf{x}_m, \mathbf{x}_n) \leq \epsilon$

Completeness:

- A space \mathcal{X} is complete if the limits of every Cauchy series are elements of \mathcal{X}
- We call $\bar{\mathcal{X}}$ the completion of \mathcal{X} , i.e. the union of \mathcal{X} and the limits of all Cauchy series in \mathcal{X}
- The real line \mathbb{R} and complex plane \mathbb{C} are complete
- The set \mathbb{Q} of rationals is not complete!

Vector Space:

A set \mathcal{X} such that $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$ and $\forall \alpha \in \mathbb{R}$ we have

- $\mathbf{x} + \mathbf{y} \in \mathcal{X}$ (**Addition**)
- $\alpha \mathbf{x} \in \mathcal{X}$ (**Multiplication**)

Examples:

- Rational numbers \mathbb{Q} over the rational field
- Real numbers \mathbb{R}
- Also true for \mathbb{R}^n

Counterexamples:

- $f : [0, 1] \rightarrow [0, 1]$ does not form a vector space!
- \mathbb{Z} is not a vector space over the real field
- The alphabet $\{a, \dots, z\}$ is not a vector space! (How do you define $+$ and \times operators?)

Normed Space:

A pair $(\mathcal{X}, \|\cdot\|)$, where \mathcal{X} is a vector space and $\|\cdot\| : \mathcal{X} \rightarrow \mathbb{R}_0^+$ is a normed space if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$ and all $\alpha \in \mathbb{R}$ it satisfies

- $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$
- $\|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\|$ (**Scaling**)
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (**Triangle inequality**)

A norm not satisfying the first condition is called a pseudo norm

Norm and Metric:

A norm induces a metric via $d(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|$

Banach Space:

A complete (in the metric defined by the norm) vector space \mathcal{X} together with a norm $\|\cdot\|$

ℓ_p^m Spaces:

Take \mathbb{R}^m endowed with the norm $\| \mathbf{x} \| := \left(\sum_{i=1}^m |x_i|^p \right)^{\frac{1}{p}}$
where $p > 0$

ℓ_p Spaces:

- These are subspaces of $\mathbb{R}^{\mathbb{N}}$ with $\| \mathbf{x} \| := \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{\frac{1}{p}}$
- The sum might not converge for all series
- For instance $x_i = \frac{1}{i}$ is in ℓ_2 but not in ℓ_1

Function Spaces $L_p(\mathcal{X})$:

- For a continuous function $f : \mathcal{X} \rightarrow \mathbb{R}$ define
 $\|f\| := \left(\int_{\mathcal{X}} |f(x)|^p dx \right)^{\frac{1}{p}}$
- Might not be well defined for all functions
- We will see more about L_2 functions later in the course

Inner Product Space:

A pair $(\mathcal{X}, \|\cdot\|)$, where \mathcal{X} is a vector space and $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_0^+$ is an inner product space if $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ and all $\alpha \in \mathbb{R}$ it satisfies

● $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$ (**Additivity**)

● $\langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$ (**Linearity**)

● $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ (**Symmetry**)

● $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \iff \mathbf{x} = 0$

Dot Product and Norm:

A dot product induces a norm via $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$

Hilbert Space:

A complete (in the metric induced by the dot product) vector space \mathcal{X} , endowed with a dot product $\langle \cdot, \cdot \rangle$

Euclidean Spaces:

Take \mathbb{R}^m endowed with the dot product $\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^m x_i y_i$

l_2 Spaces:

- Infinite series of real numbers
- We define a dot product as $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{\infty} x_i y_i$

Function Spaces $L_2(\mathcal{X})$:

- For continuous functions $f, g : \mathcal{X} \rightarrow \mathbb{C}$ define
$$\langle f, g \rangle := \int_{\mathcal{X}} \overline{f(x)} g(x) dx$$
- We take the complex conjugate of f and replace the sum by an integral

Polarization Inequality:

To recover the dot product from the norm compute
$$\| \mathbf{x} + \mathbf{y} \|^2 - \| \mathbf{x} \|^2 - \| \mathbf{y} \|^2 = 2\langle \mathbf{x}, \mathbf{y} \rangle$$

Matrix:

A real matrix $M \in \mathbb{R}^{m \times n}$ is a linear map from \mathbb{R}^m to \mathbb{R}^n

Symmetry:

- A symmetric matrix $M \in \mathbb{R}^{m \times m}$ satisfies $M_{ij} = M_{ji}$
- An anti-symmetric matrix satisfies $M_{ij} = -M_{ji}$

Range and Null Space:

For $M \in \mathbb{R}^{m \times n}$

- Its range is $\{y \in \mathbb{R}^m \mid y = Mx \text{ for some } x \in \mathbb{R}^n\}$
- Its null space is $\{x \in \mathbb{R}^n \mid Mx = 0\}$
- We have the relation $n = \dim(\text{null space}) + \dim(\text{range})$

Definition:

If $M \in \mathbb{R}^{m \times n}$, rank M is the largest number of columns of M that constitute a linearly independent set

Characteristics:

The following are equivalent for a rank k matrix M

- Exactly k rows (columns) of M are linearly independent
- Dimension of range of M is k
- \exists a k -by- k sub-matrix of M with non-zero determinant
- All $(k + 1)$ -by- $(k + 1)$ sub-matrices have determinant 0

Properties:

- For $M \in \mathbb{R}^{m \times n}$, $\text{rank}(M) \leq \min\{m, n\}$
- Deleting rows/columns can only decrease the rank
- For $M, N \in \mathbb{R}^{m \times n}$, $\text{rank}(M + N) \leq \text{rank}(M) + \text{rank}(N)$

Similar Matrices:

Two matrices $M, N \in \mathbb{R}^{m \times m}$ are similar if \exists a non-singular $S \in \mathbb{R}^{m \times m}$ such that $M = S^{-1}NS$

Eigenvalues, Eigenvectors:

Given $M \in \mathbb{R}^{m \times m}$

- An eigen pair (\mathbf{x}, λ) satisfy $M \mathbf{x} = \lambda \mathbf{x}$

Properties:

- The characteristic polynomial of M is defined as $\det(\lambda \mathbf{1} - M) = 0$
- Eigenvalues are roots of the characteristic polynomial
- Similar matrices have the same eigenvalues
- All eigenvalues of symmetric matrices are real
- If $M \in \mathbb{R}^{m \times m}$ has m distinct eigenvalues, then it is diagonalizable

Diagonalizable:

- A matrix $M \in \mathbb{R}^{m \times m}$ is diagonalizable if it is similar to a diagonal matrix
- A symmetric real matrix is always diagonalizable!

Matrix Decomposition:

We can decompose a symmetric real matrix as $O^T \Lambda O$ where O orthogonal and Λ diagonal

Orthogonality:

All eigenvectors of symmetric matrices M with different eigenvalues are mutually orthogonal

Proof For two distinct eigen pairs (\mathbf{x}, λ) and (\mathbf{x}', λ')

$$\lambda \mathbf{x}^T \mathbf{x}' = (M \mathbf{x})^T \mathbf{x}' = \mathbf{x}^T (M^T \mathbf{x}') = \mathbf{x}^T (M \mathbf{x}') = \lambda' \mathbf{x}^T \mathbf{x}'$$

hence $\lambda' = \lambda$ or $\mathbf{x}^T \mathbf{x}' = 0$ ■

Orthonormal Set:

A set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is orthonormal if $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0$ if $i \neq j$ and $\|\mathbf{x}_i\| = 1$ for all i

Orthogonal Matrix:

- An orthogonal matrix $M \in \mathbb{R}^{m \times m}$ is made up of orthonormal rows and columns
- Not difficult to see that $MM^T = \mathbf{1}$
- Equivalently $M^{-1} = M^T$

Properties:

- Orthogonal transformations preserve matrix norms

Trace:

- Trace is the sum of the diagonal elements
- For symmetric matrices $\text{tr}(MN) = \text{tr}(NM)$
- Orthogonal matrices preserve trace since $\text{tr}(O^T M O) = \text{tr}(M O O^T) = \text{tr} M$
- It can be shown that $\text{tr}(M) = \sum_{i=1}^m \lambda_i$

Determinant:

Antisymmetric multi-linear form, i.e. swapping columns or rows changes the sign, adding elements in rows and columns is linear. Useful form

$$\det M = \prod_{i=1}^m \lambda_i$$

Invariant under orthogonal transformations

Definition:

We call a function $\| \cdot \| : \mathbb{R}^{m \times m} \rightarrow \mathbb{R}_0^+$ a matrix norm if for all $M, N \in \mathbb{R}^{m \times m}$ we have

- $\|M\| = 0$ iff $M = 0$
- $\|cM\| = |c|\|M\|$ for all $c \in \mathbb{R}$
- $\|M + N\| = \|M\| + \|N\|$
- $\|MN\| = \|M\|\|N\|$ (Cauchy Schwartz?)

Matrix norms are closely related to the eigenvalues of the matrix (more on this later)

Examples:

- The ℓ_1 norm is defined as $\|M\|_1 := \sum_{i,j=1}^m |m_{ij}|$
- The ℓ_2 norm is defined as $\|M\|_1 := \left(\sum_{i,j=1}^m m_{ij}^2 \right)^{\frac{1}{2}}$

Operator Norm: Using $M \in \mathbb{R}^{m \times m}$ we have

$$\begin{aligned}\|M\|^2 &= \max_{\mathbf{x} \in \mathbb{R}^m} \frac{\|M \mathbf{x}\|^2}{\|\mathbf{x}\|^2} \\ &= \max_{\mathbf{x} \in \mathbb{R}^m \text{ and } \|\mathbf{x}\|=1} \|M \mathbf{x}\|^2 \\ &= \max_{\mathbf{x} \in \mathbb{R}^m \text{ and } \|\mathbf{x}\|=1} \mathbf{x}^\top O \Lambda O^\top O \Lambda O \mathbf{x} \\ &= \max_{\mathbf{x}' \in \mathbb{R}^m \text{ and } \|\mathbf{x}'\|=1} \mathbf{x}'^\top \Lambda^2 \mathbf{x}' \\ &= \max_{i \in [m]} \lambda_i^2.\end{aligned}$$

Frobenius Norm:

Likewise we obtain $\|M\|_{\text{Frob}}^2 = \text{tr} O \Lambda O^\top O \Lambda O^\top = \text{tr} \Lambda^2 = \sum_{i=1}^m \lambda_i^2$

Positive Definite Matrix:

A matrix $M \in \mathbb{R}^{m \times m}$ for which for all $\mathbf{x} \in \mathbb{R}^m$ we have

$$\mathbf{x}^\top M \mathbf{x} \geq 0 \text{ if } \mathbf{x} \neq 0$$

This matrix has only positive eigenvalues since for all eigenvectors \mathbf{x} we have $\mathbf{x}^\top M \mathbf{x} = \lambda \mathbf{x}^\top \mathbf{x} = \lambda \|\mathbf{x}\|^2 > 0$ and thus $\lambda > 0$.

Induced Norms and Metrics:

Every positive definite matrix induces a norm via

$$\|\mathbf{x}\|_M^2 := \mathbf{x}^\top M \mathbf{x}$$

- The triangle inequality can be seen by writing

$$\|\mathbf{x} + \mathbf{x}'\|_M^2 = (\mathbf{x} + \mathbf{x}')^\top M^{\frac{1}{2}} M^{\frac{1}{2}} (\mathbf{x} + \mathbf{x}') = \|M^{\frac{1}{2}} (\mathbf{x} + \mathbf{x}')\|^2$$

and using the triangle inequality for $M^{\frac{1}{2}} \mathbf{x}$ and $M^{\frac{1}{2}} \mathbf{x}'$.

Idea:

Can we find something similar to the eigenvalue / eigenvector decomposition for arbitrary matrices?

Decomposition:

Without loss of generality assume $m \geq n$. For $M \in \mathbb{R}^{m \times n}$ we may write M as $U\Lambda O$ where $U \in \mathbb{R}^{m \times n}$, $O \in \mathbb{R}^{n \times n}$, and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$.

Furthermore $O^T O = O O^T = U^T U = \mathbf{1}$.

Useful Trick:

Nonzero eigenvalues of $M^T M$ and $M M^T$ are the same. This is so since $M^T M \mathbf{x} = \lambda \mathbf{x}$ and hence $(M M^T) M \mathbf{x} = \lambda M \mathbf{x}$ or equivalently $(M M^T) \mathbf{x}' = \lambda \mathbf{x}'$

Basic Idea:

We have events, denoted by sets $X \subset \mathcal{X}$ in a space of possible outcomes \mathcal{X} . Then $\Pr(X)$ tells us how likely is that an event \mathbf{x} with $\mathbf{x} \in X$ will occur.

Basic Axioms:

- $\Pr(X) \in [0, 1]$ for all $X \subseteq \mathcal{X}$
- $\Pr(\mathcal{X}) = 1$
- $\Pr(\cup_i X_i) = \sum_i \Pr(X_i)$ if $X_i \cap X_j = \emptyset$ for all $i \neq j$

I am hiding gory details about σ -algebra on \mathcal{X} here

Simple Corollary:

$$\Pr(X_i \cup X_j) = \Pr(X_i) + \Pr(X_j) - \Pr(X_i \cap X_j)$$

Two Sets:

We can consider the space of events $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ and ask how likely events in the product space are

Independence:

- If the events $X \subseteq \mathcal{X}$ and $Y \subseteq \mathcal{Y}$ are independent we have $\Pr(X, Y) = \Pr(X) \cdot \Pr(Y)$
- Here $\Pr(X, Y)$ is the probability that any (\mathbf{x}, \mathbf{y}) with $\mathbf{x} \in X$ and $\mathbf{y} \in Y$ occur

Conditional Probability:

- Knowing that some event has happened will change our belief about the probability of related events i.e. $\Pr(Y|X) \Pr(X) = \Pr(Y, X)$
- This implies $\Pr(Y, X) \leq \min(\Pr(X), \Pr(Y))$

Marginalization:

We can sum out parts of a joint distribution to get the marginal distribution of a subset: $\Pr(\mathbf{x}) = \sum_{\mathbf{y}} \Pr(\mathbf{x}, \mathbf{y})$

Bayes Rule:

- Using conditional probabilities

$$\Pr(X|Y) \Pr(Y) = \Pr(X, Y) = \Pr(Y, X) = \Pr(Y|X) \Pr(X)$$

- Bayes' rule:

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

Application:

- Can infer how likely a hypothesis is, given some experimental evidence

AIDS-Test:

- We want to find out likely it is that a patient *really* has AIDS (event X) if the test is positive (event Y)
- Roughly 0.1% of all Australians are infected ($\Pr(X) = 0.001$)
- The probability of a false positive is say 1% ($\Pr(Y|\bar{X}) = 0.01$ and $\Pr(Y|X) = 1$)
- By Bayes' rule

$$\begin{aligned}\Pr(X|Y) &= \frac{\Pr(Y|X) \Pr(X)}{\Pr(Y|X) \Pr(X) + \Pr(Y|\bar{X}) \Pr(\bar{X})} \\ &= \frac{1 \times 0.001}{1 \times 0.001 + 0.01 \times 0.999} = 0.091\end{aligned}$$

- The probability of having AIDS even when the test is positive is just 9.1%!

Reliability of Eye-Witness:

- An eye-witness is 90% sure and that there were 20 people at the crime scene
- What is the probability that the guy identified committed the crime?
- Bayes' rule again

$$\Pr(X|Y) = \frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.1 \times 0.95} = 0.3213 = 32\%$$

- That's a worry ...

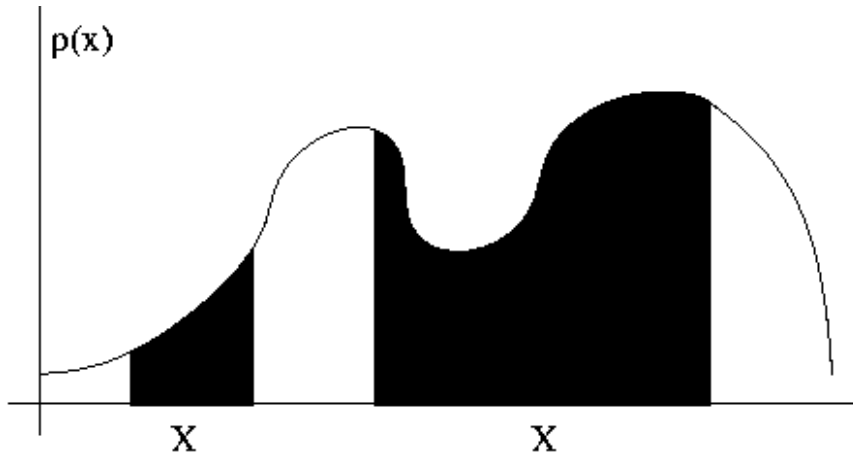
Computing $\Pr(X)$:

If we deal with continuous valued X we need integrals.

$$\Pr(X) := \int_X d\Pr(x) = \int_X p(x)dx$$

Note that the last equality only holds if such a $p(x)$ exists. For the rest of this course we assume that such a p exists

...



Multivariate Densities:

Densities on product spaces $(\mathcal{X} \times \mathcal{Y})$ are given by $p(\mathbf{x}, \mathbf{y})$

Conditional Densities:

For independent variables the densities factorize and we have

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

For dependent variables (i.e. \mathbf{x} tells us something about \mathbf{y} and vice versa) we obtain

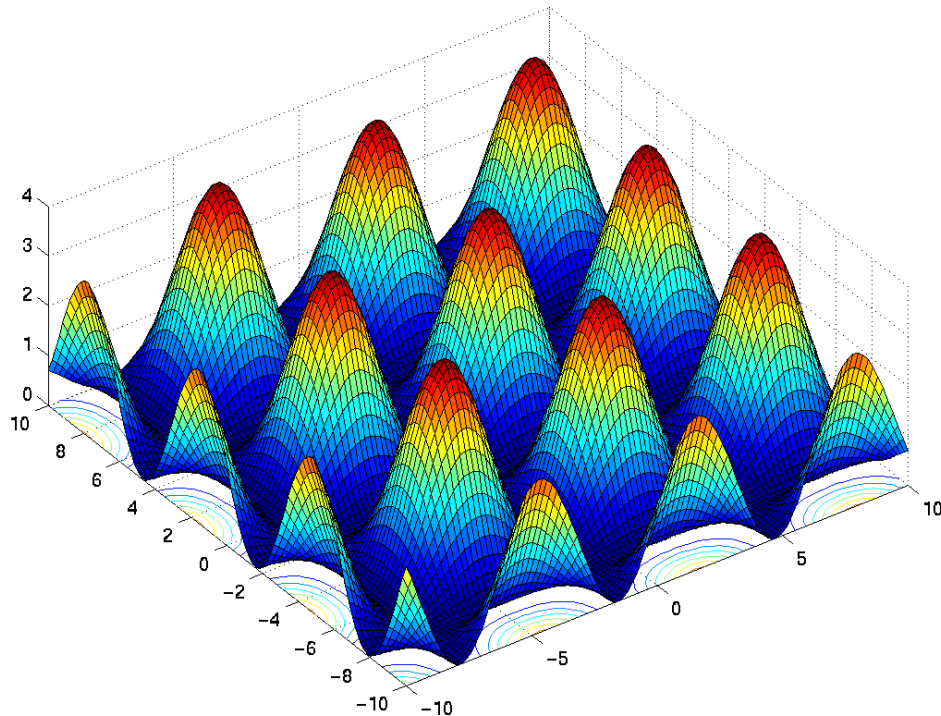
$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x} | \mathbf{y})p(\mathbf{y}) = p(\mathbf{y} | \mathbf{x})p(\mathbf{x})$$

Bayes' Rule:

Solving for $p(\mathbf{y} | \mathbf{x})$ yields

$$p(\mathbf{y} | \mathbf{x}) = \frac{p(\mathbf{x} | \mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}$$

Example: $p(\mathbf{x}) = 1 + \sin \mathbf{x}$



A factorizing distribution

Definition:

If we want to denote the fact that variables \mathbf{x} and \mathbf{y} are drawn at random from an underlying distribution, we call them *random variables*

IID variables:

- Independent and Identically Distributed RV
- Density factorizes into

$$p(\{\mathbf{x}_1, \dots, \mathbf{x}_m\}) = \prod_{i=1}^m p(\mathbf{x}_i)$$

Dependent Random Variables:

For prediction purposes we want to estimate \mathbf{y} from \mathbf{x} . In this case we *want* that \mathbf{y} is *dependent* on \mathbf{x} . If $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ we could not predict at all!

Marginalization:

Given $p(\mathbf{x}, \mathbf{y})$ we can integrate out \mathbf{y} to obtain $p(\mathbf{x})$ via

$$p(\mathbf{x}) = \int_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) d\mathbf{y}$$

Conditioning:

If we know \mathbf{y} , we can obtain $p(\mathbf{x} | \mathbf{y})$ via Bayes rule, i.e.

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

A similar trick, however, is to note that the dependence of the RHS on \mathbf{x} lies only in $p(\mathbf{x}, \mathbf{y})$ and therefore we obtain

$$p(\mathbf{x} | \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{\int_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}) d\mathbf{x}}$$

Definition:

The expectation of a function $f(\mathbf{x})$ with respect to the random variable \mathbf{x} is defined as

$$\mathbf{E}_{\mathbf{x}}[f(\mathbf{x})] := \int_{\mathbf{x}} f(\mathbf{x}) d\Pr(\mathbf{x}) = \int_{\mathbf{x}} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

The last equation is valid if a density exists

Intuition:

It is the mean value we get by sampling a large number of \mathbf{x} according to $p(\mathbf{x})$ and evaluating $f(\mathbf{x})$ on the drawn sample

Other Facts:

- Moments are expectations of higher orders
- Knowledge of all the moments completely determines the distribution

Uniform Distribution:

Assume the uniform distribution on $[0, 10]$. What is the expected value of $f(\mathbf{x}) = \mathbf{x}^2$?

$$\mathbf{E}_{\mathbf{x}}[f(\mathbf{x})] = \int_{[0,10]} f(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \int_{[0,10]} \mathbf{x}^2 \frac{1}{10}d\mathbf{x} = 33\frac{1}{3}$$

Roulette:

What is the expected loss in roulette when we bet on a number, say j (we win 36\$:1\$ if the number is hit and 0\$:1\$ otherwise)?

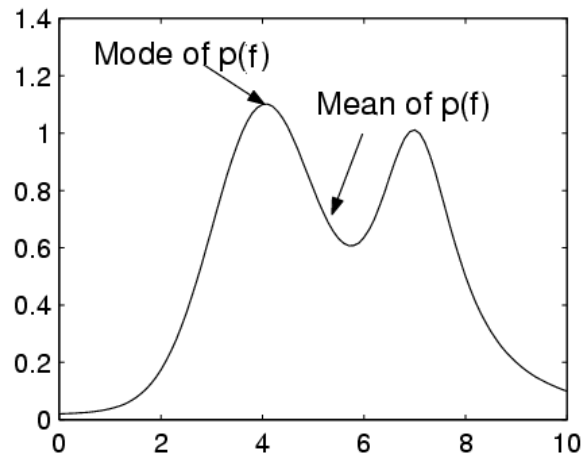
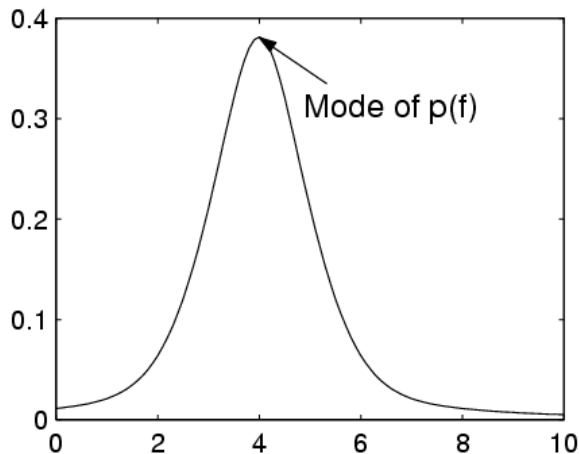
$$\mathbf{E}_{\mathbf{x}}[f(\mathbf{x})] = \sum_{i=1, i \neq j}^{37} -1\$ \cdot \frac{1}{37} + 35\$ \cdot \frac{1}{37} = -\frac{1}{37}\$$$

Mean:

- Expected value of the random variable i.e. $\mu := \mathbf{E}_x[\mathbf{x}]$

Mode:

- Largest value of the density $p(\mathbf{x})$
- Most frequently observed values of \mathbf{x}
- Mode and mean do not coincide in general



Definition:

- Amount of variation in the random variable
- First center and then compute second order moment

$$\sigma^2 := \mathbf{E}_{\mathbf{x}} \left[(\mathbf{x} - \mathbf{E}_{\mathbf{x}}[\mathbf{x}])^2 \right] = \mathbf{E}_{\mathbf{x}} \mathbf{x}^2 - (\mathbf{E}_{\mathbf{x}}[\mathbf{x}])^2$$

Normalization:

- Rescale data to zero mean and unit variance
- Preprocess data by $\mathbf{x} \rightarrow \frac{\mathbf{x} - \mu}{\sigma}$

Tails of Distributions:

- Note that the variance need not always exist
- Tails of distributions give an idea about how sharply concentrated the distribution is around its mean
- Long-tailed distributions can be killers for insurance companies!

Markov's Inequality:

If \mathbf{x} takes only non-negative values then

$$\Pr(\mathbf{x} \geq a) \leq \mathbf{E}[\mathbf{x}]/a$$

Proof:

We write

$$\begin{aligned}\mathbf{E}[\mathbf{x}] &= \int_0^a \mathbf{x} p(\mathbf{x}) d\mathbf{x} + \int_a^\infty \mathbf{x} p(\mathbf{x}) d\mathbf{x} \\ &\geq \int_a^\infty \mathbf{x} p(\mathbf{x}) d\mathbf{x} \quad \text{non-negativity} \\ &\geq a \int_a^\infty p(\mathbf{x}) d\mathbf{x} = a \Pr(\mathbf{x} \geq a)\end{aligned}$$

Observation:

Completely independent of the distribution!

Chebyshev's Inequality:

For any random variable \mathbf{x} we can bound deviations of \mathbf{x} from its mean $\mathbf{E}[\mathbf{x}]$ by

$$\Pr(|\mathbf{x} - \mathbf{E}[\mathbf{x}]| > C) \leq \frac{\sigma^2}{C^2}$$

Proof:

Apply Markov's inequality to $\mathbf{y} := (\mathbf{x} - \mathbf{E}[\mathbf{x}])^2$

Applications:

- Information about some measurement is easy to get
- Easy to estimate the variance too
- Don't know anything about the distribution :-)
- Still can make statements about probability of deviating from the mean!

Jensen's Inequality

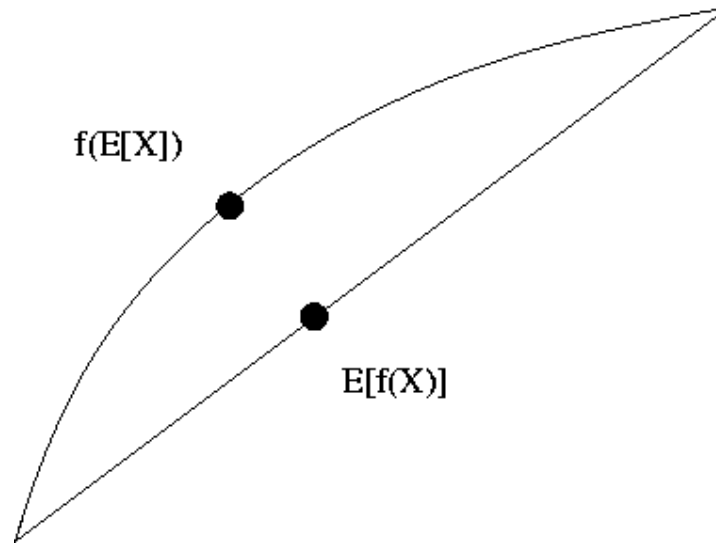
Jensen's Inequality:

If f is a convex function and X is a random variable then

$$\mathbf{E}[f(X)] \geq f(\mathbf{E}[X])$$

Picture:

Notice how expectation is a linear operator



Definition:

- Measures the *disorder* of a system
- Defined as

$$H(p) = - \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x})$$

Properties:

- $H(p) > 0$ unless only one possible outcome
- Maximal value occurs for uniform p
- Deep connections to information theory exist

The Formula:

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

Mean:

- Notice that $p(\mu + \xi) = p(\mu - \xi)$ because $((\mu + \xi) - \mu)^2 = \xi^2 = ((\mu - \xi) - \mu)^2$
- Hence the mean of $p(x)$ is μ

Variance:

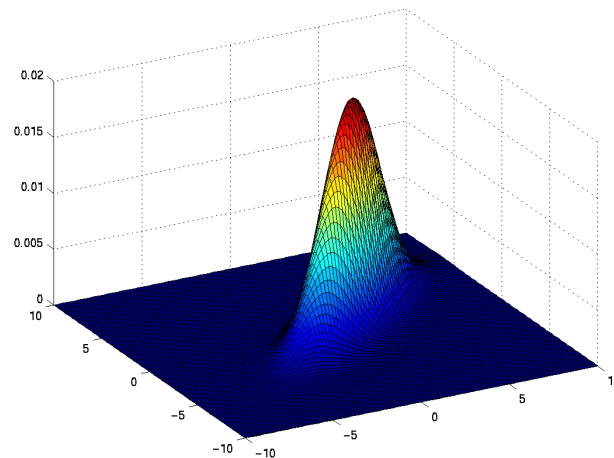
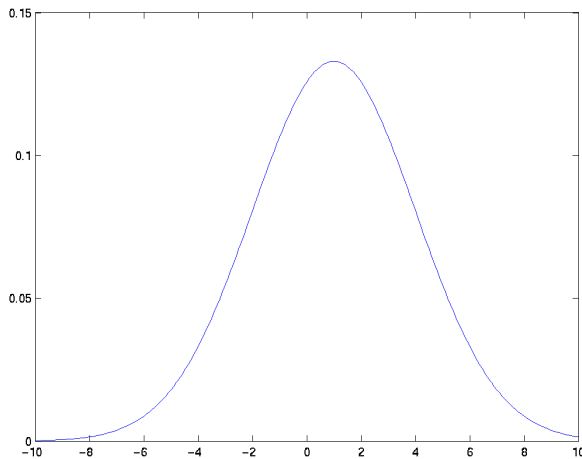
The variance of $p(x)$ is σ^2 . We show this by proving that

$$\begin{aligned}\text{Var}x &= \int_{\mathbb{R}} p(x)(x - \mu)^2 dx = \int_{\mathbb{R}} p(\mu + \xi)\xi^2 d\xi \\ &= \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{\xi^2}{2}} \xi^2 d\xi = \sigma^2\end{aligned}$$

Pictures of Normal Distributions

Normal Distribution in \mathbb{R} : Mean 1, Variance 3

Normal Distribution in \mathbb{R}^2 : Mean $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, Variance $\begin{bmatrix} 6 & 4 \\ 4 & 4 \end{bmatrix}$



Covariance:

- For a multivariate distribution

$$\text{Cov } \mathbf{x} := \mathbf{E} [(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]$$

- We now compute a matrix instead of a single number
- In particular

$$(\text{Cov } \mathbf{x})_{ij} = \mathbf{E} [(x_i - \mu_i)(x_j - \mu_j)]$$

Correlated Variables:

- Measures degree of association between variables
- If positively correlated then

$$(\text{Cov } \mathbf{x})_{ij} = \sqrt{(\text{Cov } \mathbf{x})_{ii} (\text{Cov } \mathbf{x})_{jj}}$$

- For uncorrelated variables their covariance vanishes

The Formula:

- $\Sigma \in \mathbb{R}^{m \times m}$ is positive definite
- The mean $\mu \in \mathbb{R}^m$

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m \det \Sigma}} \exp \left(-\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu) \right)$$

Mean:

Obviously this is μ (we can check that by symmetry)

Variance:

- Tedious calculation shows that $\text{Var}(\mathbf{x}) = \Sigma$
- Hint: decompose $\Sigma = O^\top \Lambda O$
- Hint: use $\det \Sigma = \det O \Sigma O^\top = \prod_i \lambda_i$, where λ_i are the eigenvalues of Σ

Decay of Atoms:

- The probability that a atom decays within 1 sec is $1 - p$
- The probability that it decays in n sec is $1 - p^n$
- In the continuous domain the probability of decay after time T is

$$P(\xi \leq T) = 1 - \exp(-\lambda T) = \int_0^T p(t) dt$$

Laplacian Distribution:

- Consequently, $p(t)$ is given by $\lambda \exp(-\lambda T)$
- It is a particularly long-tailed distribution

Mean and Variance:

- Mean is given by $\mu = \frac{1}{\lambda}$
- Variance is given by $\frac{2}{\lambda^2}$

Why Gaussians are good for you: If we have many independent errors, the net effect will be a single error with normal distribution.

Theorem:

Denote by ξ_i random variables with variance $\sigma_i \leq \bar{\sigma}$ for some $\bar{\sigma}$ and with mean $\mu_i \leq \bar{\mu}$ for some μ , then the random variable $\xi := \frac{\sum_{i=1}^m \xi_i - \mu_i}{\sqrt{\sum_{i=1}^m \sigma_i^2}}$ has zero mean and unit variance.

Furthermore for $m \rightarrow \infty$ the random variable ξ will be normally distributed.

Sum of Random Variables:

- Consider the average of m random variables $\xi_i \in [0, 1]$

$$\xi := \frac{1}{m} \sum_{i=1}^m \xi_i$$

- Will ξ be concentrated around its mean?

Hoeffding's Theorem:

- For any $\varepsilon > 0$ the probability of large deviations of ξ from $\mathbf{E}[\xi]$ is bounded by

$$\Pr (|\xi - \mathbf{E}[\xi]| \geq \varepsilon) \leq 2 \exp (-2 \varepsilon^2 m)$$

- things get exponentially better, the more random variables we average over (i.e. more the number of observations)

Questions?