Introduction to Machine Learning

THE AUSTRALIAN NATIONAL UNIVERSITY

What you can use it for

- pattern recognition (faces, digits, speech),
- bioinformatics (gene finding, introns)
- internet (spam filtering, search engines)
- prediction (stock market)

What you get

- skills in programming, numerical analysis, optimization
- practical experience with data
- easy do-it-yourself algorithms

http://axiom.anu.edu.au/∼smola/engn4520/

Overview

THE AUSTRALIAN NATIONAL UNIVERSITY

Week 1

Linear Algebra, Hilbert Spaces, Numerical Mathematics

Week 2

Learning Theory, Statistics, Risk Functional, Common Distributions, Perceptron

Week 3

Regression, Squared Loss, Noise Models and Loss, Regularization

Week 4

Kernels, Kernel Perceptron, Kernel Regression

Week 5

Large Margin and Optimization, SV Classification, Regression, Novelty Detection

Week 6

Applications and Mini Projects: Text Categorization and Bad Digits

Practical Issues

Scoring

This is a 3 credit point unit. Exercises and programming each count $\frac{1}{4}$, the final exam counts $\frac{1}{2}$.

Problem Sheets

Due Monday at 10am in the mailbox. **Late submissions cost 20% a day**.

You are expected to work together in groups of 3 and submit **one solution sheet per group**. If you copy from other groups you will not get points for these solutions.

Tutorials

Ben O'Loghlin (`ben@syseng.anu.edu.au`) will hold the tutorials (Thursday 2-5pm) which include solutions of the exercise sheets and some programming practice with the SVLab toolbox.

Final Exam

Probably Monday, June 18 (slides, personal notes, calculator and tables are OK).

A Crash-Course in Math

THE AUSTRALIAN NATIONAL UNIVERSITY

Topics

- Vector spaces, Hilbert and Banach Spaces, Metrics and Norms
- Matrices, Eigenvalues, Orthogonal Transformations, Singular Values
- Operators, Operator Norms, Function Spaces revisited

Rationale

- We need this toolbox to describe the functions we will be dealing with and to set up the optimization/learning problems.

Metric

Definition 1 (Metric)

Denote by $\mathcal{X}$ *a space. Then* $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$ *is a metric on* $\mathcal{X}$ *if for all* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$

1. $d(\mathbf{x}, \mathbf{y}) = 0$ *is equivalent to* $\mathbf{x} = \mathbf{y}$
2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ *(Triangle Inequality)*

Example 1 (Trivial Metric)

For arbitrary $\mathcal{X}$ *define* $d(\mathbf{x}, \mathbf{y}) = 1$ *if* $\mathbf{x} \neq \mathbf{y}$ *and* $d(\mathbf{x}, \mathbf{y}) = 0$ *if* $\mathbf{x} = \mathbf{y}$.

Example 2 (Manhattan Distance)

For $\mathcal{X} = \mathbb{R}^n$ *define* $d(\mathbf{x}, \mathbf{y}) := \sum_{i=1}^{n} |x_i - y_i|$.

THE AUSTRALIAN NATIONAL UNIVERSITY

Definition 2 (Vector Space)

A space $\mathcal{X}$ on which for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and for all $\alpha \in \mathbb{R}$ the following operations are defined:

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

Definition 3 (Cauchy Series)

Given a space $\mathcal{X}$, a series $\mathbf{x}_i \in \mathcal{X}$ with $i \in \mathbb{N}$ is a Cauchy series if for any ϵ there exists an n_0 such that for all $m, n \geq n_0$ we have $d(\mathbf{x}_m, \mathbf{x}_n) \leq \varepsilon$.

Definition 4 (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 all its limits of Cauchy series.

Vector Spaces: Examples

Rational Numbers

Addition and multiplication are obviously OK. However, the space **is not complete**. For instance, we can find a Caucy series of $x_i \in \mathbb{Q}$ converging to $\sqrt{2}$.

Real Numbers

Addition and multiplication are obviously OK. The same holds for limits (recall algebra lectures).

$\mathbb{R}^n$

Prototypical example of a vector space. addition, multiplication, and limits are obviously OK, e.g., take $\mathcal{X} = \mathbb{R}^5$ and $\mathbf{x} = (2, 33.4, 4.2, 2.999, 6)$.

Polynomials

Functions such as $f(x) := a + bx + cx^2 + dx^3$ obviously form a vector space. For polynomials of finite order n we can even find a mapping between $\mathcal{X}$ and $\mathbb{R}^n$.

Vector Spaces: Examples

THE AUSTRALIAN NATIONAL UNIVERSITY

Series

series (a_i) of numbers with $a_i \in \mathbb{R}$ and $i \in \mathbb{N}$ are clearly vector spaces.

Fourier Expansions

expansions via the discrete Fourier transform form a vector space where

$$f(x) = \sum_{j=1}^{n} s_j \sin(jx) + c_j \cos(jx)$$

Functions

many classes of functions, e.g., $f : [0, 1] \to \mathbb{R}$.

Counterexamples

- $f : [0, 1] \to [0, 1]$ does not form a vector space!
- $\mathbb{Z}$ is not a vector space, unless we only allow multiplications by integers.
- The alphabet $\{\mathrm{a}, \ldots, \mathrm{z}\}$ is not a vectorspace (still it can be an interesting mathematical object, e.g. when determining similarity of documents).

Banach Spaces

THE AUSTRALIAN NATIONAL UNIVERSITY

Definition 5 (Norm)
Given a vector space $\mathcal{X}$, a mapping $\|\cdot\| : \mathcal{X} \to \mathbb{R}_0^+$ is called a norm if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and all $\alpha \in \mathbb{R}$ it satisfies

1. $\|\mathbf{x}\| = 0$ *if and only if* $\mathbf{x} = 0$
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ *(scaling)*
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ *(triangle inequality)*

A mapping $\|\cdot\|$ not satisfying (1) is called **pseudo norm**.
Note that a norm also introduces a **metric** via $d(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|$.

Definition 6 (Banach Space)
A complete vector space $\mathcal{X}$ together with a norm $\|\cdot\|$.

Banach Spaces: Examples

THE AUSTRALIAN NATIONAL UNIVERSITY

ℓ_p^m Spaces

Take the $\mathbb{R}^m$ endowed with the norm $\|\mathbf{x}\| := \left(\sum_{i=1}^m |x_i|^p\right)^{\frac{1}{p}}$ where $p > 0$. Note that in $\mathbb{R}^m$ all norms are **equivalent**, i.e. there exist $c, C \in \mathbb{R}^+$ such that

$$c\|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq C\|\mathbf{x}\|_a \text{ for all } \mathbf{x} \in \mathcal{X} \text{ and likewise } \frac{1}{C}\|\mathbf{x}\|_b \leq \|\mathbf{x}\|_a \leq \frac{1}{c}\|\mathbf{x}\|_b$$

ℓ_p Spaces

These are subspaces of $\mathbb{R}^{\mathbb{N}}$ with $\|\mathbf{x}\| := \left(\sum_{i=1}^\infty |x_i|^p\right)^{\frac{1}{p}}$.

Not for all series x_i the sum converges, e.g., $x_i = \frac{1}{i}$ is in ℓ_2 but not in ℓ_1.

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

We replace sums by integrals over $\mathcal{X}$ and obtain $\|f\| := \left(\int_{\mathcal{X}} |f(x)|^p dx\right)^{\frac{1}{p}}$. Again, not for all f this integral is defined, i.e. they are not elements of the corresponding $L_p(\mathcal{X})$.

Hilbert Spaces

Definition 7 (Dot Product)

Given a vector space $\mathcal{X}$, a mapping $\langle \cdot, \cdot \rangle$ with $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which for all $\alpha \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ satisfies

1. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ *(symmetry)*
2. $\langle \mathbf{x}, \alpha \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$ *(linearity)*
3. $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$ *(additivity)*

Definition 8 (Hilbert Space)

A complete vector space $\mathcal{X}$, endowed with a dot product $\langle \cdot, \cdot \rangle$.

The dot product automatically generates a norm (and a metric) by $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. Thus Hilbert spaces are special case of a Banach space.

These are the spaces we will work with in this lecture.

Hilbert Spaces: Examples

Euclidean Spaces Use standard dot product for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ given by $\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^{m} x_i y_i$

Function Spaces ($L_2(X)$) Functions on X with $f : X \to \mathbb{C}$ for all $f \in \mathfrak{F}$. Here we can define the dot product for $f, g \in \mathfrak{F}$ by $\langle f, g \rangle := \int_X \overline{f(x)} g(x) dx$ Note that we take the complex conjugate of f. Also note that all we did was to replace the sum by an integral.

ℓ_2 (Infinite) series of real numbers, $\ell_2 \subset \mathbb{R}^{\mathbb{N}}$. We define a dot product for $\mathbf{x}, \mathbf{y} \in \ell_2$ by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{\infty} x_i y_i$

Polarization Inequality We can recover the dot product from the norm $\|\mathbf{x}\|$ by computing $\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2 = 2\langle \mathbf{x}, \mathbf{y} \rangle$.

Matrices

THE AUSTRALIAN NATIONAL UNIVERSITY

In the following we assume that a matrix $M \in \mathbb{R}^{m \times n}$ corresponds to a linear map from $\mathbb{R}^m$ to $\mathbb{R}^n$ and is given by its entries $M_{ij} \in \mathbb{R}$.

Symmetry

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

Antisymmetry

An antisymmetric matrix $M \in \mathbb{R}^{m \times m}$ satisfies $M_{ij} = -M_{ji}$.

Rank

Denote by I the image of $\mathbb{R}^m$ under $M \in \mathbb{R}^{m \times n}$. Since M is a linear map, we can find a I as a linear combination of vectors. $\operatorname{rank}(M)$ is the smallest number of such vectors that span I.

Orthogonality

A matrix $M \in \mathbb{R}^{m \times m}$ with $M^\top M = \mathbf{1}$ is called an orthogonal matrix (if $M \in \mathbb{C}^{m \times m}$ it is called unitary). This means $M^\top = M^{-1}$.

THE AUSTRALIAN NATIONAL UNIVERSITY

Orthogonality, Part II

It consists of mutually orthogonal rows and columns. The corresponding matrix group is often denoted by $\mathrm{O}(m)$ (the orthogonal group). If it is only a rotation, it is called $\mathrm{SO}(m)$ (special orthogonal group).

Note that from $M^\top M = \mathbf{1}$ it also follows that $MM^\top = \mathbf{1}$ since $M^\top M = \mathbf{1} \Rightarrow (MM^\top)M = M$ (and all matrices have full rank).

Example

Rotation matrices in $\mathbb{R}^2$ are given by

$$M = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} \text{ here } \det M = 1.$$

Mirror matrices are

$$M = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \text{ and } M = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \text{ here } \det M = 1.$$

Matrix Invariants

THE AUSTRALIAN NATIONAL UNIVERSITY

Trace:

$\mathrm{tr} M := \sum_{i=1}^{m} M_{ii}$ One can show $\mathrm{tr}(AB) = \mathrm{tr}(BA)$and thus for symmetric matrices

$$\mathrm{tr} M = \mathrm{tr}(O^\top \Lambda O) = \mathrm{tr}(\Lambda O O^\top) = \mathrm{tr} \Lambda = \sum_{i=1}^{m} \lambda_i$$

Determinant:

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

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

Both trace and determinant are invariant under orthogonal transformations $M \to O^\top M O$ where $O \in \mathrm{SO}(m)$ for of the matrix.

Matrix Norms

Operator Norm

The norm of a linear operator A between two Banach spaces $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$\|A\| := \max_{\mathbf{x} \in \mathcal{X}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$$

This clearly satisfies all conditions of a norm:

- $\|\alpha A\| = \max_{\mathbf{x} \in \mathcal{X}} \frac{\|\alpha A\mathbf{x}\|}{\|\mathbf{x}\|} = |\alpha| \|A\|$.
- $\|A + B\| = \max_{\mathbf{x} \in \mathcal{X}} \frac{\|(A+B)\mathbf{x}\|}{\|\mathbf{x}\|} \leq \max_{\mathbf{x} \in \mathcal{X}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} + \max_{\mathbf{x} \in \mathcal{X}} \frac{\|B\mathbf{x}\|}{\|\mathbf{x}\|} = \|A\| + \|B\|$
- $\|A\| = 0$ implies $\max_{\mathbf{x} \in \mathcal{X}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = 0$ and thus $A\mathbf{x} = 0$ for all $\mathbf{x}$. This means that $A = 0$.

Frobenius Norm

For a matrix $M \in \mathbb{R}^{m \times n}$ we can define a norm in analogy to the vector norm by

$$\|M\|^2_{\text{Frob}} = \sum_{i=1}^{m} \sum_{j=1}^{n} M_{ij}^2$$

Definition 9 (Eigenvalues, Eigenvectors)
Denote by $M \in \mathbb{R}^{m \times m}$ matrix, then an eigenvalue $\lambda \in \mathbb{R}$ and eigenvector $\mathbf{x} \in \mathbb{R}^m$ satisfy

$$M\mathbf{x} = \lambda\mathbf{x}$$

Ananlogously for operators $A : \mathcal{X} \to \mathcal{X}$ we have $A\mathbf{x} = \lambda\mathbf{x}$.

Caveat

We cannot always find a complete eigensystem. Example: $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$

Symmetric Matrices

All eigenvalues of symmetric matrices are real and symmetric matrices are fully diagonalizable, i.e. we can find m eigenvectors.

Eigensystems

THE AUSTRALIAN NATIONAL UNIVERSITY

Orthogonality:

All eigenvectors of symmetric matrices M with different eigenvalues are mutually orthogonal. Proof: for two eigenvectors $\mathbf{x}$ and $\mathbf{x}'$ with eigenalues λ, λ' use

$$\lambda \mathbf{x}^\top \mathbf{x}' = (M\mathbf{x})^\top \mathbf{x} = \mathbf{x}^\top (M^\top \mathbf{x}') = \mathbf{x}^\top (M\mathbf{x}') = \lambda' \mathbf{x}^\top \mathbf{x}' \text{ hence } \lambda' = \lambda \text{ or } \mathbf{x}^\top \mathbf{x}' = 0.$$

Matrix Decomposition:

We can decompose symmetric $M \in \mathbb{R}^{m \times m}$ into $O^\top \Lambda O$ where $O \in SO(n)$ and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$.

Example:

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \text{ has eigenvalues } (-1, 3) \text{ and eigenvectors } v_1 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, v_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

Matrix Norms Revisited

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

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

Frobenius Norm: Likewise we obtain

$$\|M\|_{\text{Frob}}^2 = \operatorname{tr}(MM^\top) = \operatorname{tr} O \Lambda O^\top O \Lambda O^\top = \operatorname{tr} \Lambda O^\top O \Lambda O^\top O = \operatorname{tr} \Lambda^2 = \sum_{i=1}^m \lambda_i^2$$

Positive Matrices

THE AUSTRALIAN NATIONAL UNIVERSITY

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 Metrices:

Every positive definite matrix induces a norm via

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

- Linearity is obvious, so is uniqueness
- 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 $M^{\frac{1}{2}}\mathbf{x}'$.

Singular Value Decompositions

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 $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

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

Useful Trick:

Nonzero eigenvalues of $M^\top M$ and $M M^\top$ are the same. This is so since

$$M^\top M \mathbf{x} = \lambda \mathbf{x} \text{ and hence } (M M^\top) M \mathbf{x} = \lambda M \mathbf{x} \text{ or equivalently } (M M^\top) \mathbf{x}' = \lambda \mathbf{x}'.$$