Modern Algorithms for Machine Learning

Lecture 1 (Part II): Nearest Neighbour and Dimensionality Reduction

Thomas Sauerwald

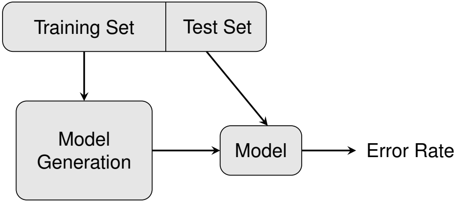
University of Cambridge, Department of Computer Science
email: thomas.sauerwald@cl.cam.ac.uk

July 2020

UNIVERSITY OF CAMBRIDGE

Outline

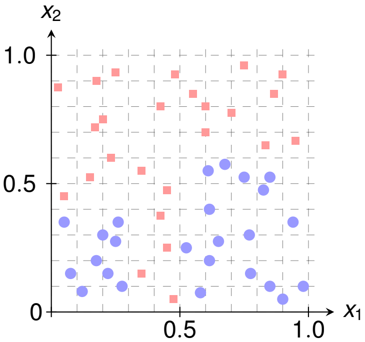
Machine learning model (Supervised Learning)



- Training data: used to build a model or classifier for the data.
- Test data: helps validate it.
- Overfitting:
  - ML algorithms tune their model to the training set (and not the general data that the training set represents);
  - test data helps restricting overfitting.
- Feature selection: which to use to input into ML algorithm?
- Training set creation: where do labels for data come from?

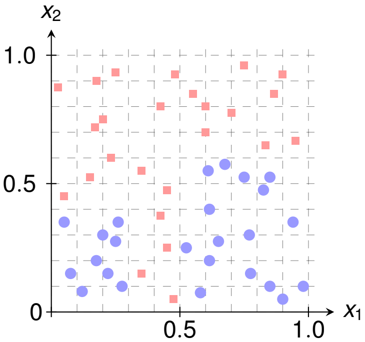
The Model (A bit more formal...)

- Let $\mathcal{X} = \mathbb{R}^d$ (domain set)
- Let $\mathcal{Y} = \{-1, +1\}$ (label set)
- Let $S = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m))$ (training set)



The Model (A bit more formal...)

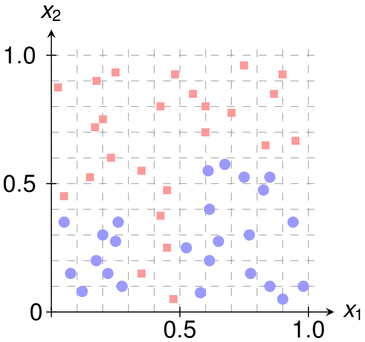
- Let $\mathcal{X} = \mathbb{R}^d$ (domain set)
- Let $\mathcal{Y} = \{-1, +1\}$ (label set)
- Let $S = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m))$ (training set)



Goal: Find a classifier $h : \mathcal{X} \to \mathcal{Y}$, which labels any unseen data point.

The Model (A bit more formal...)

- Let $\mathcal{X} = \mathbb{R}^d$ (domain set)
- Let $\mathcal{Y} = \{-1, +1\}$ (label set)
- Let $S = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m))$ (training set)

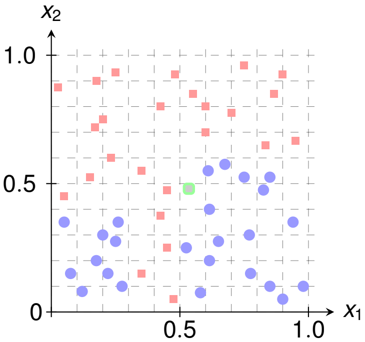


Goal: Find a classifier $h : \mathcal{X} \to \mathcal{Y}$, which labels any unseen data point.

sometimes also called hypothesis or prediction rule

The Model (A bit more formal...)

- Let $\mathcal{X} = \mathbb{R}^d$ (domain set)
- Let $\mathcal{Y} = \{-1, +1\}$ (label set)
- Let $S = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m))$ (training set)



Goal: Find a classifier $h : \mathcal{X} \to \mathcal{Y}$, which labels any unseen data point.

sometimes also called hypothesis or prediction rule

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$

Towards the *k*-NN: A Simple Example in One Dimension

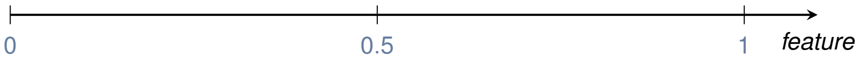
Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??

Towards the *k*-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



Towards the *k*-NN: A Simple Example in One Dimension

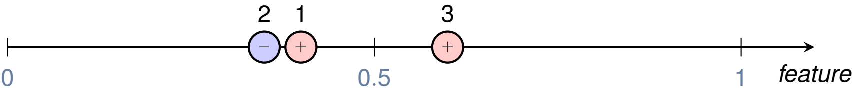
Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??

Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??

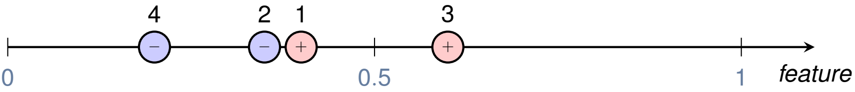
Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



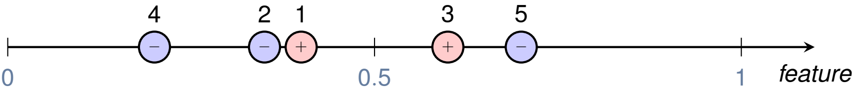
Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



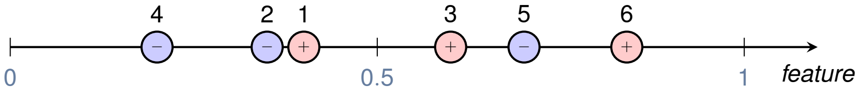
Towards the *k*-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



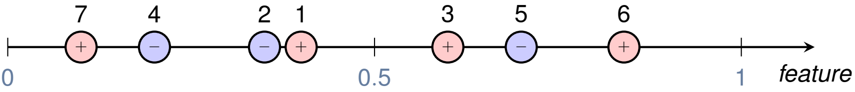
Towards the *k*-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



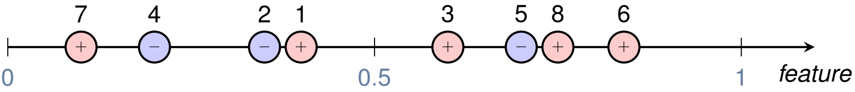
Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



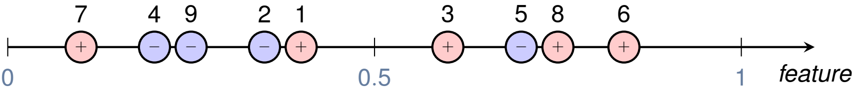
Towards the *k*-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



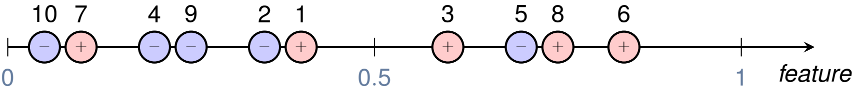
Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



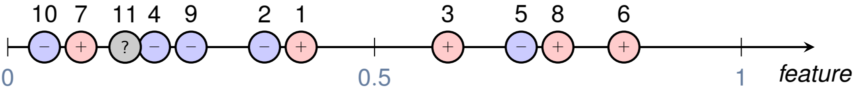
Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



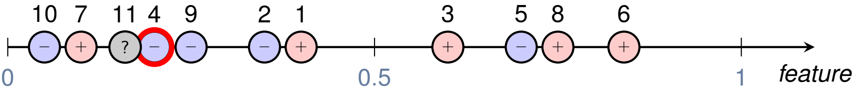
Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



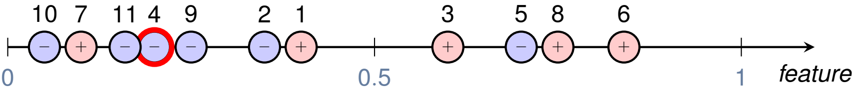
Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



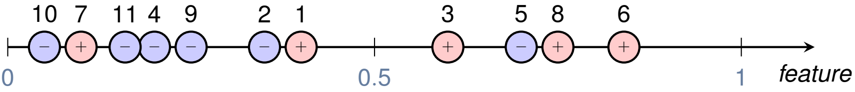
Towards the *k*-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	??



Towards the k-NN: A Simple Example in One Dimension

Input	Feature	Label
Data Point 1	0.4	$+$
Data Point 2	0.35	$-$
Data Point 3	0.6	$+$
Data Point 4	0.2	$-$
Data Point 5	0.7	$-$
Data Point 6	0.84	$+$
Data Point 7	0.1	$+$
Data Point 8	0.75	$+$
Data Point 9	0.25	$-$
Data Point 10	0.05	$-$
Test Point 11	0.16	$-$



The k-NN Algorithm

- Let $d(\mathbf{x}, \mathbf{x}') := \sqrt{\sum_{i=1}^{d} (x_i - x_i')^2}$ be the Euclidean Distance

The k-NN Algorithm

- Let $d(\mathbf{x}, \mathbf{x}') := \sqrt{\sum_{i=1}^{d} (x_i - x_i')^2}$ be the Euclidean Distance
- For every $\mathbf{x} \in \mathcal{X}$, let $\pi_1(\mathbf{x}), \ldots, \pi_m(\mathbf{x})$ be a reordering of $\{1, 2, \ldots, m\}$ according to their distance to $\mathbf{x}$, i.e., for every $1 \leq i < m$:

$$\rho(\mathbf{x}, x_{\pi_i(\mathbf{x})}) \leq \rho(\mathbf{x}, x_{\pi_{i+1}(\mathbf{x})}).$$

The k-NN Algorithm

- Let $d(\mathbf{x}, \mathbf{x}') := \sqrt{\sum_{i=1}^{d}(x_i - x_i')^2}$ be the Euclidean Distance
- For every $\mathbf{x} \in \mathcal{X}$, let $\pi_1(\mathbf{x}), \ldots, \pi_m(\mathbf{x})$ be a reordering of $\{1, 2, \ldots, m\}$ according to their distance to $\mathbf{x}$, i.e., for every $1 \leq i < m$:

$$\rho(\mathbf{x}, x_{\pi_i(\mathbf{x})}) \leq \rho(\mathbf{x}, x_{\pi_{i+1}(\mathbf{x})}).$$

For each point $\mathbf{x}$, we sort the elements in training set increasingly in their distance to $\mathbf{x}$.

The k-NN Algorithm

- Let $d(\mathbf{x}, \mathbf{x}') := \sqrt{\sum_{i=1}^{d}(x_i - x_i')^2}$ be the Euclidean Distance
- For every $\mathbf{x} \in \mathcal{X}$, let $\pi_1(\mathbf{x}), \ldots, \pi_m(\mathbf{x})$ be a reordering of $\{1, 2, \ldots, m\}$ according to their distance to $\mathbf{x}$, i.e., for every $1 \le i < m$:

$$\rho(\mathbf{x}, x_{\pi_i(\mathbf{x})}) \le \rho(\mathbf{x}, x_{\pi_{i+1}(\mathbf{x})}).$$

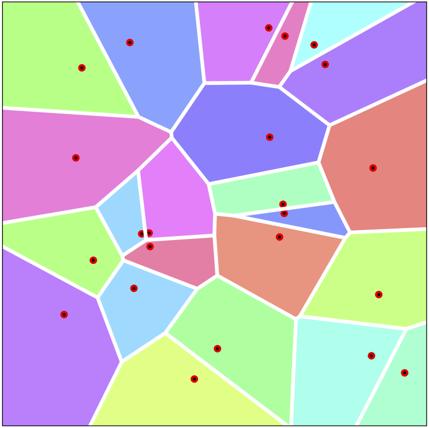
For each point $\mathbf{x}$, we sort the elements in training set increasingly in their distance to $\mathbf{x}$.

k-NN

input: a training sample $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
output: for every point $\mathbf{x} \in \mathcal{X}$,
return the majority label among $\{y_{\pi_i(\mathbf{x})} : i \le k\}$

Source: SS&BD

Special Case: 1-NN



- For $k = 1$, the produced decision boundaries are Voronoi-cells
- Any new point will be classified according to the centre of each cell

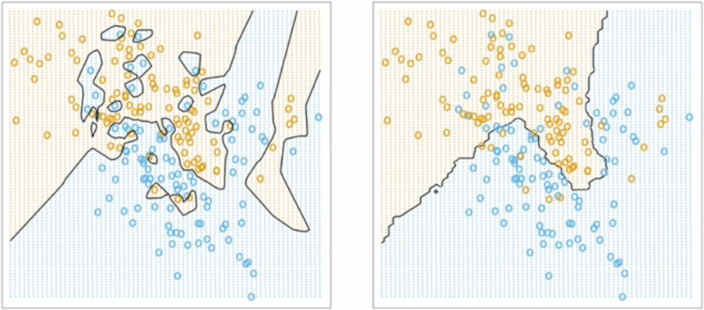
How many Neighbours should we choose?

Source: Lecture by Ulrike von Luxburg

$k = 1$ $k = 15$

How many Neighbours should we choose?

For small k, k-NN overfits the data!

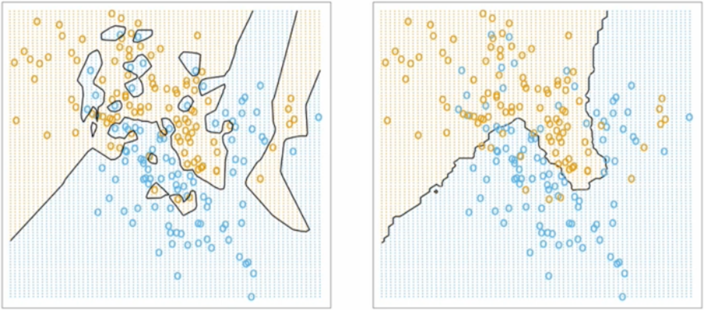


Source: Lecture by Ulrike von Luxburg

$k = 1$ $\qquad$ $k = 15$

How many Neighbours should we choose?

For small k, k-NN overfits the data!



Source: Lecture by Ulrike von Luxburg

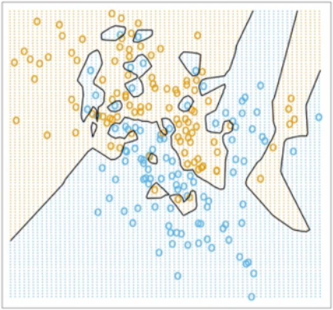
$k = 1$ $\qquad\qquad$ $k = 15$

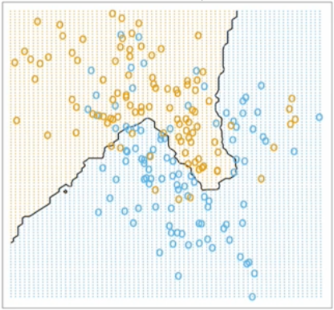
Question: What happens if $k = n$, where n is the total number of points in our training set?

How many Neighbours should we choose?

For small k, k-NN overfits the data!

Rule-of-thumb: $k \approx \log n$ is good





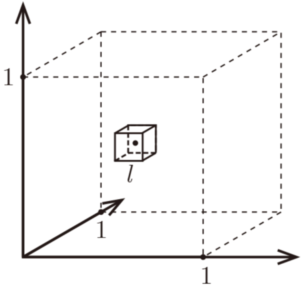
Source: Lecture by Ulrike von Luxburg

$k = 1$ $\qquad$ $k = 15$

Question: What happens if $k = n$, where n is the total number of points in our training set?

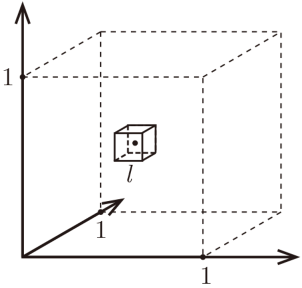
???

Curse of Dimensionality



Source: Kilian Weinberger

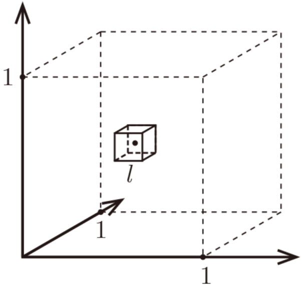
Curse of Dimensionality



Source: Kilian Weinberger

- Suppose $n = 1000$ points are "randomly" spread across $[0, 1]^d$

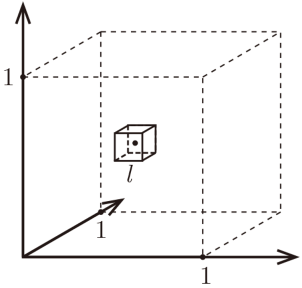
Curse of Dimensionality



Source: Kilian Weinberger

- Suppose $n = 1000$ points are "randomly" spread across $[0, 1]^d$
- If we want to find the 10 nearest neighbour, how large must be the subcube be?

Curse of Dimensionality

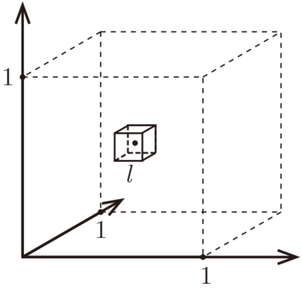


Source: Kilian Weinberger

- Suppose $n = 1000$ points are "randomly" spread across $[0, 1]^d$
- If we want to find the 10 nearest neighbour, how large must be the subcube be?

d	ℓ
2	0.1
10	0.63
100	0.955
1000	0.9954

Curse of Dimensionality



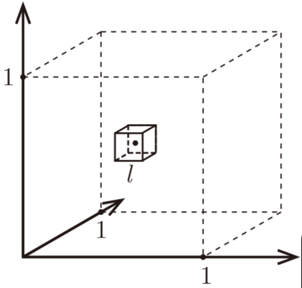
Source: Kilian Weinberger

- Suppose $n = 1000$ points are "randomly" spread across $[0, 1]^d$
- If we want to find the 10 nearest neighbour, how large must be the subcube be?

d	ℓ
2	0.1
10	0.63
100	0.955
1000	0.9954

To find the closest neighbour, we need to search the entire space!

Curse of Dimensionality



Source: Kilian Weinberger

In high dimensions, almost all points have the same (far) distance!

- Suppose $n = 1000$ points are “randomly” spread across $[0, 1]^d$
- If we want to find the 10 nearest neighbour, how large must be the subcube be?

d	ℓ
2	0.1
10	0.63
100	0.955
1000	0.9954

To find the closest neighbour, we need to search the entire space!

Summary of Nearest Neighbour

+ very easy to implement and understand
+ does not use any training/learning phase
+ can be applied to almost any prediction problem
  (often a good “baseline”)

Summary of Nearest Neighbour

+ very easy to implement and understand
+ does not use any training/learning phase
+ can be applied to almost any prediction problem (often a good “baseline”)
– results may crucially depend on the choice of distance function and k
– finding nearest neighbour is costly ($\rightsquigarrow$ KD-trees, locality-sensitive hashing)
– very slow with large k or high-dimensional data (course of dimensionality)

We will now learn a powerful pre-processing method called **Dimensionality Reduction**!

Outline

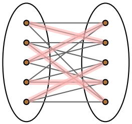
Mathematical Tools

- **Matrices and Geometry**
  - Data points (predictions, observations, classifications) encoded in matrices/vectors
  - This allows geometric representation that is the basis of many ML methods
  - Networks and graphs $\Leftrightarrow$ adjacency matrices

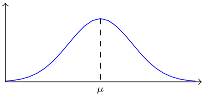
  Inner product, Hyperplanes, Eigenvectors

- **Probability Theory**
  - Sampling-based algorithms
  - Algorithm exploits concentration of measure
  - Incomplete data, noisy/corrupted data can be modelled by a random processes

  Random Variables, Chernoff Bounds, hashing



$$A = \begin{pmatrix} 0&0&0&0&0&1&1&0&0&1\\ 0&0&0&0&0&0&1&1&0&1\\ 0&0&0&0&0&1&1&0&1&0\\ 0&0&0&0&0&1&0&1&1&0\\ 0&0&0&0&0&0&0&1&1&1\\ 1&0&1&1&0&0&0&0&0&0\\ 1&1&1&0&0&0&0&0&0&0\\ 0&1&0&1&1&0&0&0&0&0\\ 0&0&1&1&1&0&0&0&0&0\\ 1&1&0&0&1&0&0&0&0&0 \end{pmatrix}$$



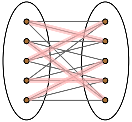
Mathematical Tools

- **Matrices and Geometry**
  - Data points (predictions, observations, classifications) encoded in matrices/vectors
  - This allows geometric representation that is the basis of many ML methods
  - Networks and graphs ⇔ adjacency matrices

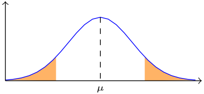
  Inner product, Hyperplanes, Eigenvectors

- **Probability Theory**
  - Sampling-based algorithms
  - Algorithm exploits concentration of measure
  - Incomplete data, noisy/corrupted data can be modelled by a random processes

  Random Variables, Chernoff Bounds, hashing

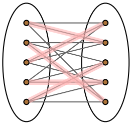


$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$



- **Matrices and Geometry**
  - Data points (predictions, observations, classifications) encoded in matrices/vectors
  - This allows geometric representation that is the basis of many ML methods
  - Networks and graphs $\Leftrightarrow$ adjacency matrices

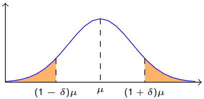
Inner product, Hyperplanes, Eigenvectors



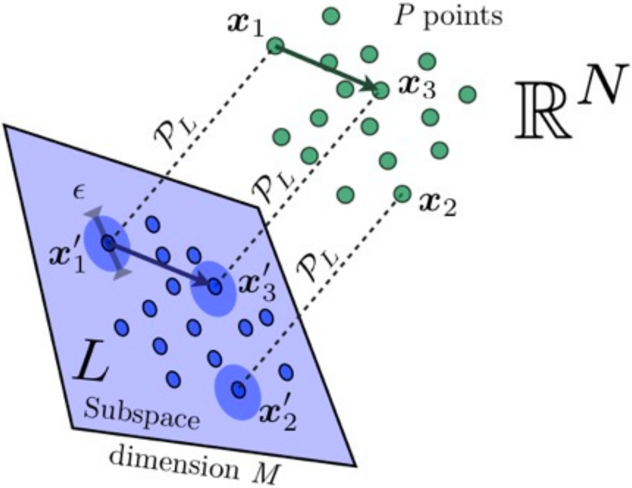
$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

- **Probability Theory**
  - Sampling-based algorithms
  - Algorithm exploits concentration of measure
  - Incomplete data, noisy/corrupted data can be modelled by a random processes

Random Variables, Chernoff Bounds, hashing

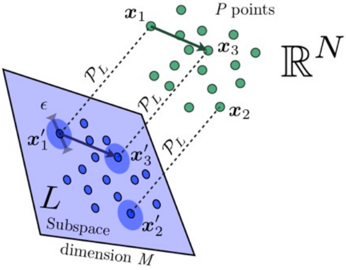


Intuition behind Dimensionality Reduction

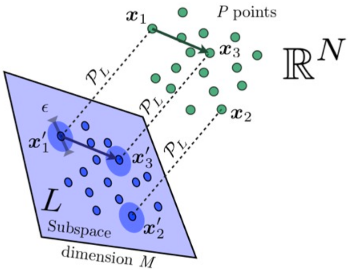


Source: Laurent Jacques

Dimensionality Reduction: Basic Setup

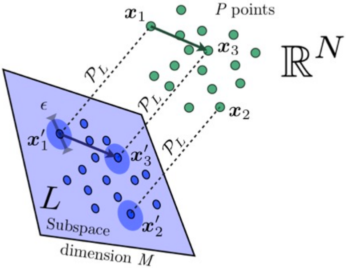


Dimensionality Reduction: Basic Setup



- Given P points $x_1, x_2, \ldots, x_P \in \mathbb{R}^N$
- Want to find P points $x'_1, x'_2, \ldots, x'_P \in \mathbb{R}^M$, $M \ll N$

Dimensionality Reduction: Basic Setup

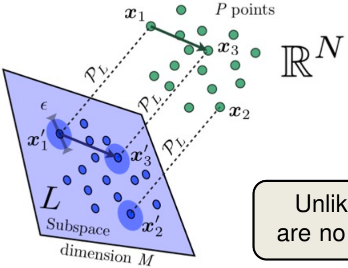


- Given P points $x_1, x_2, \ldots, x_P \in \mathbb{R}^N$
- Want to find P points $x'_1, x'_2, \ldots, x'_P \in \mathbb{R}^M$, $M \ll N$

Goal: Distances are approximately preserved, i.e.,

$$(1-\epsilon) \cdot \|x_i - x_j\| \leq \|x'_i - x'_j\| \leq (1+\epsilon) \cdot \|x_i - x_j\| \qquad \text{for all } i, j$$

Dimensionality Reduction: Basic Setup



Unlike other methods like PCA, there are no assumptions on the original data.

- Given P points $x_1, x_2, \ldots, x_P \in \mathbb{R}^N$
- Want to find P points $x'_1, x'_2, \ldots, x'_P \in \mathbb{R}^M$, $M \ll N$

Goal: Distances are approximately preserved, i.e.,

$$(1-\epsilon)\cdot\|x_i - x_j\| \leq \|x'_i - x'_j\| \leq (1+\epsilon)\cdot\|x_i - x_j\| \qquad \text{for all } i, j$$

Johnson-Lindenstrauss-Lemma

Theorem

Let $x_1, x_2, \ldots, x_p \in \mathbb{R}^N$ be arbitrary. Pick any $\epsilon = (0, 1)$. Then for some $M = O(\log(P)/\epsilon^2)$, there is a polynomial-time algorithm that, with probability at least $1 - \frac{2}{P}$, computes $x_1', x_2', \ldots, x_P' \in \mathbb{R}^M$ such that

$$(1-\epsilon)\cdot\|x_i - x_j\| \leq \|x_i' - x_j'\| \leq (1+\epsilon)\cdot\|x_i - x_j\| \qquad \text{for all } i, j$$
$$(1-\epsilon)\cdot\|x_i\| \leq \|x_i'\| \leq (1+\epsilon)\cdot\|x_i\| \qquad \text{for all } i.$$

Johnson-Lindenstrauss-Lemma

Note: M does not depend on N!

Theorem

Let $x_1, x_2, \ldots, x_p \in \mathbb{R}^N$ be arbitrary. Pick any $\epsilon = (0,1)$. Then for some $M = O(\log(P)/\epsilon^2)$, there is a polynomial-time algorithm that, with probability at least $1 - \frac{2}{P}$, computes $x'_1, x'_2, \ldots, x'_P \in \mathbb{R}^M$ such that

$$(1-\epsilon) \cdot \|x_i - x_j\| \leq \|x'_i - x'_j\| \leq (1+\epsilon) \cdot \|x_i - x_j\| \qquad \text{for all } i, j$$

$$(1-\epsilon) \cdot \|x_i\| \leq \|x'_i\| \leq (1+\epsilon) \cdot \|x_i\| \qquad \text{for all } i.$$

Johnson-Lindenstrauss-Lemma

Note: M does not depend on N!

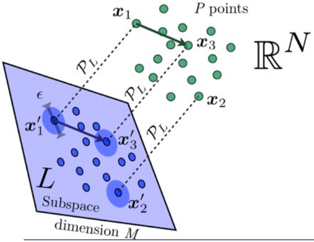
Theorem

Let $x_1, x_2, \ldots, x_p \in \mathbb{R}^N$ be arbitrary. Pick any $\epsilon = (0, 1)$. Then for some $M = O(\log(P)/\epsilon^2)$, there is a polynomial-time algorithm that, with probability at least $1 - \frac{2}{P}$, computes $x'_1, x'_2, \ldots, x'_P \in \mathbb{R}^M$ such that

$$(1-\epsilon) \cdot \|x_i - x_j\| \leq \|x'_i - x'_j\| \leq (1+\epsilon) \cdot \|x_i - x_j\| \qquad \text{for all } i, j$$
$$(1-\epsilon) \cdot \|x_i\| \leq \|x'_i\| \leq (1+\epsilon) \cdot \|x_i\| \qquad \text{for all } i.$$

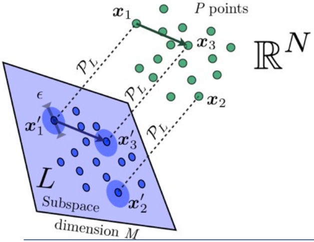
How to construct $x'_1, x'_2, \ldots, x'_P$?

Key Tool: Random Projection Method



Key Tool: Random Projection Method

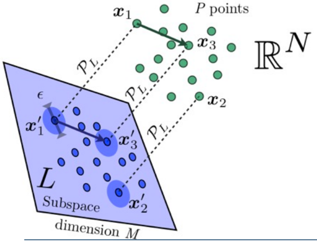
Definition of $f : \mathbb{R}^N \to \mathbb{R}^M \qquad (M \ll N)$



Key Tool: Random Projection Method

$$\text{Definition of } f : \mathbb{R}^N \to \mathbb{R}^M \qquad (M \ll N)$$

$$f \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_N \end{pmatrix} = \begin{pmatrix} \cdots\cdots r_1^T \cdots\cdots \\ \cdots\cdots r_2^T \cdots\cdots \\ \vdots \\ \cdots\cdots r_M^T \cdots\cdots \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_N \end{pmatrix}$$

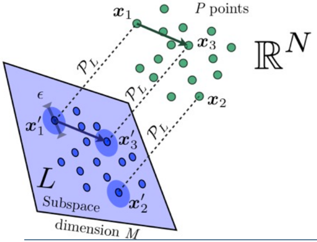


Key Tool: Random Projection Method

Definition of $f : \mathbb{R}^N \to \mathbb{R}^M \qquad (M \ll N)$

$$f\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_N \end{pmatrix} = \begin{pmatrix} \cdots\cdots r_1^T \cdots\cdots \\ \cdots\cdots r_2^T \cdots\cdots \\ \vdots \\ \cdots\cdots r_M^T \cdots\cdots \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_N \end{pmatrix} = \begin{pmatrix} r_1^T v \\ r_2^T v \\ \vdots \\ r_M^T v \end{pmatrix}$$

, where r_i's are random vectors



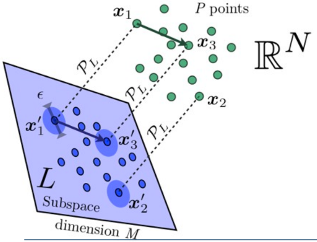
Key Tool: Random Projection Method

Definition of $f : \mathbb{R}^N \to \mathbb{R}^M \qquad (M \ll N)$

$$f\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_N \end{pmatrix} = \begin{pmatrix} \cdots\cdots r_1^T \cdots\cdots \\ \cdots\cdots r_2^T \cdots\cdots \\ \vdots \\ \cdots\cdots r_M^T \cdots\cdots \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_N \end{pmatrix} = \begin{pmatrix} r_1^T v \\ r_2^T v \\ \vdots \\ r_M^T v \end{pmatrix}$$

, where r_i's are random vectors

Each entry of r_i is independently drawn from $\mathcal{N}(0, 1)$



Key Tool: Random Projection Method

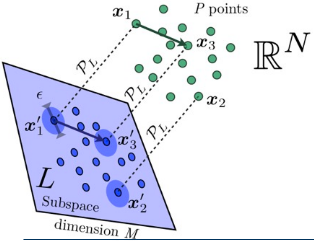
Definition of $f : \mathbb{R}^N \to \mathbb{R}^M \qquad (M \ll N)$

$$f\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_N \end{pmatrix} = \begin{pmatrix} \cdots\cdots r_1^T \cdots\cdots \\ \cdots\cdots r_2^T \cdots\cdots \\ \vdots \\ \cdots\cdots r_M^T \cdots\cdots \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_N \end{pmatrix} = \begin{pmatrix} r_1^T v \\ r_2^T v \\ \vdots \\ r_M^T v \end{pmatrix}$$

, where r_i's are random vectors

Each entry of r_i is independently drawn from $\mathcal{N}(0,1)$

r_i's are chosen independently



Key Tool: Random Projection Method

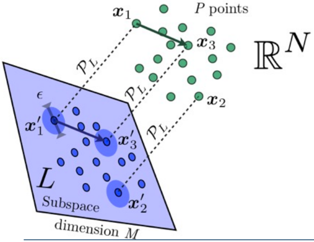
Definition of $f : \mathbb{R}^N \to \mathbb{R}^M \qquad (M \ll N)$

$$f\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_N \end{pmatrix} = \begin{pmatrix} \cdots\cdots r_1^T \cdots\cdots \\ \cdots\cdots r_2^T \cdots\cdots \\ \vdots \\ \cdots\cdots r_M^T \cdots\cdots \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_N \end{pmatrix} = \begin{pmatrix} r_1^T v \\ r_2^T v \\ \vdots \\ r_M^T v \end{pmatrix}$$

, where r_i's are random vectors

Each entry of r_i is independently drawn from $\mathcal{N}(0,1)$

r_i's are chosen independently



Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

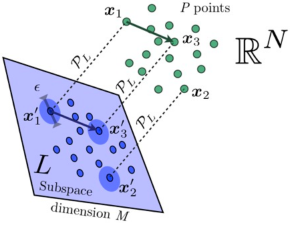
$$\mathbf{Pr}\left[1 - \epsilon \le \frac{\|f(w)\|}{\sqrt{M}} \le 1 + \epsilon\right] \ge 1 - \frac{2}{P^3}.$$

Proof of Theorem (using JL-Lemma)

Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\textbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$



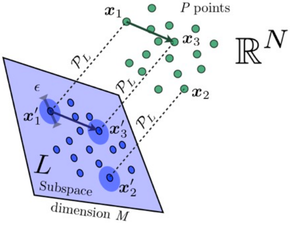
Proof of Theorem (using JL-Lemma)

Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$

- Define $L(v) := \frac{f(v)}{\sqrt{M}}$



Proof of Theorem (using JL-Lemma)

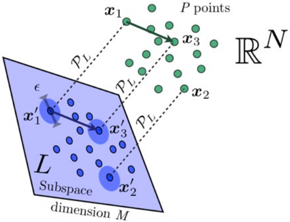
Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$

- Define $L(v) := \frac{f(v)}{\sqrt{M}}$
- JL-Lemma with $w = \frac{v}{\|v\|} \Rightarrow$

$$\mathbf{Pr}\left[\,(1-\epsilon)\cdot\|v\| \leq \|L(v)\| \leq (1+\epsilon)\cdot\|v\|\,\right] \geq 1 - \frac{2}{P^3}.$$



Proof of Theorem (using JL-Lemma)

Johnson-Lindenstrauss Lemma

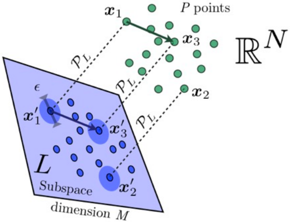
Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$

- Define $L(v) := \frac{f(v)}{\sqrt{M}}$
- JL-Lemma with $w = \frac{v}{\|v\|} \Rightarrow$

$$\mathbf{Pr}\left[(1-\epsilon)\cdot\|v\| \leq \|L(v)\| \leq (1+\epsilon)\cdot\|v\|\right] \geq 1 - \frac{2}{P^3}.$$

- Apply to $v = x_j$ and $v = x_i - x_j, j \neq i$ and the Union bound ($\mathbf{Pr}[A \cup B] \leq \mathbf{Pr}[A] + \mathbf{Pr}[B]$): W.p. $1 - \frac{2}{P}$,



$$(1-\epsilon)\cdot\|x_i - x_j\| \leq \|L(x_i - x_j)\| \leq (1+\epsilon)\cdot\|x_i - x_j\| \qquad \text{for all } i,j$$

$$(1-\epsilon)\cdot\|x_i\| \leq \|L(x_i)\| \leq (1+\epsilon)\cdot\|x_i\| \qquad \text{for all } i.$$

Proof of Theorem (using JL-Lemma)

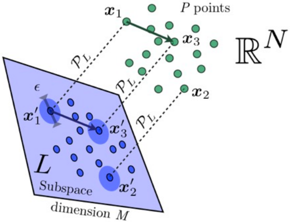
Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$

- Define $L(v) := \frac{f(v)}{\sqrt{M}}$
- JL-Lemma with $w = \frac{v}{\|v\|} \Rightarrow$

$$\mathbf{Pr}\left[\,(1-\epsilon)\cdot\|v\| \leq \|L(v)\| \leq (1+\epsilon)\cdot\|v\|\,\right] \geq 1 - \frac{2}{P^3}.$$



- Apply to $v = x_j$ and $v = x_i - x_j, j \neq i$ and the Union bound ($\mathbf{Pr}[A \cup B] \leq \mathbf{Pr}[A] + \mathbf{Pr}[B]$): W.p. $1 - \frac{2}{P}$,

$L(x_i - x_j) = L(x_i) - L(x_j)$

$$(1-\epsilon)\cdot\|x_i - x_j\| \leq \|L(x_i - x_j)\| \leq (1+\epsilon)\cdot\|x_i - x_j\| \qquad \text{for all } i,j$$

$$(1-\epsilon)\cdot\|x_i\| \leq \|L(x_i)\| \leq (1+\epsilon)\cdot\|x_i\| \qquad \text{for all } i.$$

Proof of Theorem (using JL-Lemma)

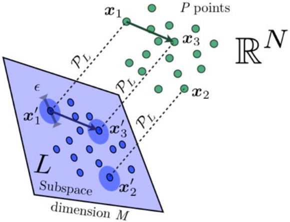
Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \le \frac{\|f(w)\|}{\sqrt{M}} \le 1 + \epsilon\right] \ge 1 - \frac{2}{P^3}.$$

- Define $L(v) := \frac{f(v)}{\sqrt{M}}$
- JL-Lemma with $w = \frac{v}{\|v\|} \Rightarrow$

$$\mathbf{Pr}\left[(1-\epsilon)\cdot\|v\| \le \|L(v)\| \le (1+\epsilon)\cdot\|v\|\right] \ge 1 - \frac{2}{P^3}.$$



- Apply to $v = x_j$ and $v = x_i - x_j, j \neq i$ and the Union bound ($\mathbf{Pr}[A \cup B] \le \mathbf{Pr}[A] + \mathbf{Pr}[B]$): W.p. $1 - \frac{2}{P}$,

$L(x_i - x_j) = L(x_i) - L(x_j)$

$$(1-\epsilon)\cdot\|x_i - x_j\| \le \|L(x_i - x_j)\| \le (1+\epsilon)\cdot\|x_i - x_j\| \quad \text{for all } i, j$$
$$(1-\epsilon)\cdot\|x_i\| \le \|L(x_i)\| \le (1+\epsilon)\cdot\|x_i\| \quad \text{for all } i. \qquad \square$$

Outline

Proof of JL-Lemma (1/4)

Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$

Proof of JL-Lemma (1/4)

Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$

Proof (of the upper bound):

Proof of JL-Lemma (1/4)

Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$

Proof (of the upper bound):

- Squaring yields $\mathbf{Pr}\left[\|f(w)\|^2 > (1+\epsilon)^2 \cdot M\right]$.

Proof of JL-Lemma (1/4)

Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$

Proof (of the upper bound):

- Squaring yields $\mathbf{Pr}\left[\,\|f(w)\|^2 > (1+\epsilon)^2 \cdot M\,\right]$.
- Recall that the i-th coordinate of $f(w)$ is $r_i^T \cdot w$. The distribution is

$$\mathcal{N}(0, \sum_{j=1}^{N} w_j^2) = \mathcal{N}(0, 1).$$

Proof of JL-Lemma (1/4)

Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$

Proof (of the upper bound):

- Squaring yields $\mathbf{Pr}\left[\|f(w)\|^2 > (1+\epsilon)^2 \cdot M\right]$.
- Recall that the i-th coordinate of $f(w)$ is $r_i^T \cdot w$. The distribution is

$$\mathcal{N}(0, \sum_{j=1}^{N} w_j^2) = \mathcal{N}(0, 1).$$

If $X_1, \ldots, X_N$ are independent random variables with distribution $\mathcal{N}(0,1)$ each, then $\sum_{j=1}^N w_j X_j$ has distribution $\mathcal{N}(0, \sum_{j=1}^N w_j^2)$

Proof of JL-Lemma (1/4)

Johnson-Lindenstrauss Lemma

Let $w \in \mathbb{R}^N$ with $\|w\| = 1$. Then for some $M = O(\log(P)/\epsilon^2)$, we have

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{\|f(w)\|}{\sqrt{M}} \leq 1 + \epsilon\right] \geq 1 - \frac{2}{P^3}.$$

Proof (of the upper bound):

- Squaring yields $\mathbf{Pr}\left[\, \|f(w)\|^2 > (1+\epsilon)^2 \cdot M \,\right]$.
- Recall that the i-th coordinate of $f(w)$ is $r_i^T \cdot w$. The distribution is

$$\mathcal{N}(0, \sum_{j=1}^{N} w_j^2) = \mathcal{N}(0,1).$$

If $X_1, \ldots, X_N$ are independent random variables with distribution $\mathcal{N}(0,1)$ each, then $\sum_{j=1}^N w_j X_j$ has distribution $\mathcal{N}(0, \sum_{j=1}^N w_j^2)$

- Hence

$$\|f(w)\|^2 = \sum_{i=1}^{M} X_i^2,$$

where the X_i's are independent $\mathcal{N}(0,1)$ random variables.

Proof of JL-Lemma (2/4)

- Taking expectations:

$$\mathbf{E}\left[\,\|f(w)\|^2\,\right] = \mathbf{E}\left[\sum_{i=1}^{M} X_i^2\right] = \sum_{i=1}^{M} \mathbf{E}\left[\,X_i^2\,\right] = M$$

Proof of JL-Lemma (2/4)

- Taking expectations:

$$\mathbf{E}\left[\|f(w)\|^2\right] = \mathbf{E}\left[\sum_{i=1}^{M} X_i^2\right]$$
$$= \sum_{i=1}^{M} \mathbf{E}\left[X_i^2\right] = M$$

- We will now derive a Chernoff bound for $X := \sum_{i=1}^{M} X_i^2$. Let $t \in (0, 1/2)$,

$$\mathbf{Pr}\left[X > \alpha\right] = \mathbf{Pr}\left[e^{tY} > e^{t\alpha}\right] \leq e^{-t\alpha} \cdot \mathbf{E}\left[e^{tX}\right].$$

Proof of JL-Lemma (2/4)

- Taking expectations:

$$\mathbf{E}\left[\|f(w)\|^2\right] = \mathbf{E}\left[\sum_{i=1}^{M} X_i^2\right]$$
$$= \sum_{i=1}^{M} \mathbf{E}\left[X_i^2\right] = M$$

- We will now derive a Chernoff bound for $X := \sum_{i=1}^{M} X_i^2$. Let $t \in (0, 1/2)$,

$$\mathbf{Pr}\left[X > \alpha\right] = \mathbf{Pr}\left[e^{tY} > e^{t\alpha}\right] \leq e^{-t\alpha} \cdot \mathbf{E}\left[e^{tX}\right].$$

- Since $X_1^2, \dots, X_M^2$ are independent,

$$\mathbf{E}\left[e^{tX}\right] = \mathbf{E}\left[e^{t\sum_{i=1}^{M} X_i^2}\right] = \mathbf{E}\left[\prod_{i=1}^{M} e^{tX_i^2}\right] \stackrel{!}{=} \prod_{i=1}^{M} \mathbf{E}\left[e^{(tX_i^2)}\right]$$

Proof of JL-Lemma (3/4)

- We need to analyse $\mathbf{E}\left[e^{tX_i^2}\right]$:

$$\mathbf{E}\left[e^{tX_i^2}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(ty^2) \exp(-y^2/2) dy$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-y^2(1-2t)/2\right) dy$$

Proof of JL-Lemma (3/4)

- We need to analyse $\mathbf{E}\left[e^{tX_i^2}\right]$:

$$\mathbf{E}\left[e^{tX_i^2}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(ty^2) \exp(-y^2/2) dy$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-y^2(1-2t)/2\right) dy$$

- Now substitute $z = y \cdot \sqrt{1-2t}$ to obtain

$$= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{1-2t}} \cdot \int_{-\infty}^{\infty} e^{-z^2/2} dz$$
$$= \frac{1}{\sqrt{1-2t}}$$

Proof of JL-Lemma (3/4)

- We need to analyse $\mathbf{E}\left[e^{tX_i^2}\right]$:

$$\mathbf{E}\left[e^{tX_i^2}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(ty^2)\exp(-y^2/2)dy$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-y^2(1-2t)/2\right) dy$$

- Now substitute $z = y \cdot \sqrt{1-2t}$ to obtain

$$= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{1-2t}} \cdot \int_{-\infty}^{\infty} e^{-z^2/2} dz$$
$$= \frac{1}{\sqrt{1-2t}}$$

$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2} dz$ is the CDF of $\mathcal{N}(0,1)$

Proof of JL-Lemma (4/4)

- Hence with $\alpha = (1+\epsilon)^2 M$,

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \leq e^{-t(1+\epsilon)^2 M} \cdot \left(\frac{1}{1-2t}\right)^{M/2}$$

Proof of JL-Lemma (4/4)

- Hence with $\alpha = (1+\epsilon)^2 M$,

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \le e^{-t(1+\epsilon)^2 M} \cdot \left(\frac{1}{1-2t}\right)^{M/2}$$

- We choose $t = (1 - 1/(1+\epsilon)^2)/2$, giving

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \le e^{(M - M(1+\epsilon)^2)/2} \cdot (1+\epsilon)^{-M}$$

Proof of JL-Lemma (4/4)

- Hence with $\alpha = (1+\epsilon)^2 M$,

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \leq e^{-t(1+\epsilon)^2 M} \cdot \left(\frac{1}{1-2t}\right)^{M/2}$$

- We choose $t = (1 - 1/(1+\epsilon)^2)/2$, giving

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \leq e^{(M - M(1+\epsilon)^2)/2} \cdot (1+\epsilon)^{-M}$$

- The last term can be rewritten as

$$\begin{aligned} &\exp\left(\frac{M}{2}\left(1-(1+\epsilon)^2\right) - \frac{M}{2}\ln\left(\frac{1}{(1+\epsilon)^2}\right)\right) \\ &= \exp\left(-M\left(\epsilon + \epsilon^2/2 - \ln(1+\epsilon)\right)\right) \end{aligned}$$

Proof of JL-Lemma (4/4)

- Hence with $\alpha = (1+\epsilon)^2 M$,

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \leq e^{-t(1+\epsilon)^2 M} \cdot \left(\frac{1}{1-2t}\right)^{M/2}$$

- We choose $t = (1 - 1/(1+\epsilon)^2)/2$, giving

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \leq e^{(M - M(1+\epsilon)^2)/2} \cdot (1+\epsilon)^{-M}$$

- The last term can be rewritten as

$$\begin{aligned} &\exp\left(\frac{M}{2}\left(1-(1+\epsilon)^2\right) - \frac{M}{2}\ln\left(\frac{1}{(1+\epsilon)^2}\right)\right) \\ &= \exp\left(-M\left(\epsilon + \epsilon^2/2 - \ln(1+\epsilon)\right)\right) \end{aligned}$$

- Using $\ln(1+x) \leq x$ for $x \geq 0$, implies

$$\begin{aligned} \mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] &\leq \exp\left(-M\left(\epsilon + \epsilon^2/2 - \epsilon\right)\right) \\ &\leq \exp\left(-M\epsilon^2/2\right). \end{aligned}$$

Proof of JL-Lemma (4/4)

- Hence with $\alpha = (1+\epsilon)^2 M$,

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \leq e^{-t(1+\epsilon)^2 M} \cdot \left(\frac{1}{1-2t}\right)^{M/2}$$

- We choose $t = (1 - 1/(1+\epsilon)^2)/2$, giving

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \leq e^{(M - M(1+\epsilon)^2)/2} \cdot (1+\epsilon)^{-M}$$

- The last term can be rewritten as

$$\begin{aligned} &\exp\left(\frac{M}{2}\left(1-(1+\epsilon)^2\right) - \frac{M}{2}\ln\left(\frac{1}{(1+\epsilon)^2}\right)\right) \\ &= \exp\left(-M\left(\epsilon + \epsilon^2/2 - \ln(1+\epsilon)\right)\right) \end{aligned}$$

- Using $\ln(1+x) \leq x$ for $x \geq 0$, implies

$$\begin{aligned} \mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] &\leq \exp\left(-M\left(\epsilon + \epsilon^2/2 - \epsilon\right)\right) \\ &\leq \exp\left(-M\epsilon^2/2\right). \end{aligned}$$

- With $M = 6 \ln P/\epsilon^2$, the last term becomes $\frac{2}{P^3}$.

Proof of JL-Lemma (4/4)

- Hence with $\alpha = (1+\epsilon)^2 M$,

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \leq e^{-t(1+\epsilon)^2 M} \cdot \left(\frac{1}{1-2t}\right)^{M/2}$$

- We choose $t = (1 - 1/(1+\epsilon)^2)/2$, giving

$$\mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] \leq e^{(M - M(1+\epsilon)^2)/2} \cdot (1+\epsilon)^{-M}$$

- The last term can be rewritten as

$$\begin{aligned} &\exp\left(\frac{M}{2}\left(1-(1+\epsilon)^2\right) - \frac{M}{2}\ln\left(\frac{1}{(1+\epsilon)^2}\right)\right) \\ &= \exp\left(-M\left(\epsilon + \epsilon^2/2 - \ln(1+\epsilon)\right)\right) \end{aligned}$$

- Using $\ln(1+x) \leq x$ for $x \geq 0$, implies

$$\begin{aligned} \mathbf{Pr}\left[X > (1+\epsilon)^2 M\right] &\leq \exp\left(-M\left(\epsilon + \epsilon^2/2 - \epsilon\right)\right) \\ &\leq \exp\left(-M\epsilon^2/2\right). \end{aligned}$$

- With $M = 6 \ln P/\epsilon^2$, the last term becomes $\frac{2}{P^3}$.
- Lower bound is derived similarly $\Rightarrow$ proof complete $\square$

Example: Target Dimension M of JL-Lemma

Recall: $M \leq \frac{6 \ln P}{\epsilon^2}$

ϵ	Number of Points P	Target Dimension M
1/2	1,000	166
1/2	10,000	221
1/2	100,000	276
1/2	1,000,000	331
1/2	10,000,000	387
1/10	1,000	4145
1/10	10,000	5526
1/10	100,000	6907
1/10	1,000,000	8298
1/10	10,000,000	9670

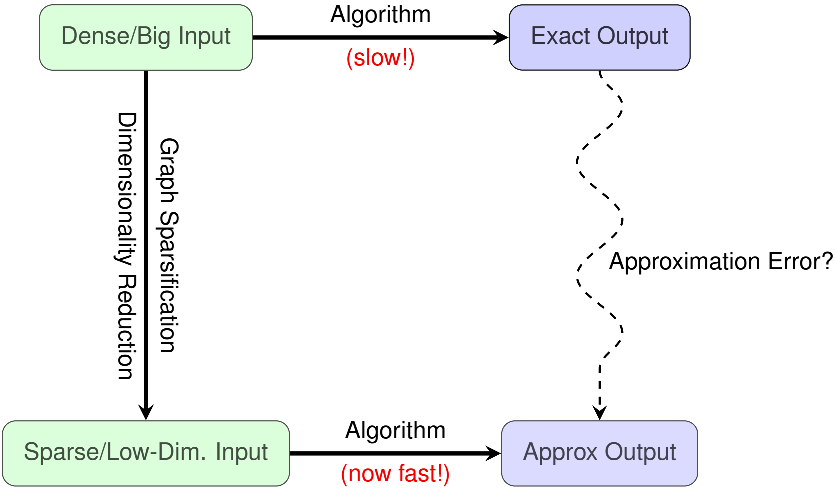
General Comments on the JL-Lemma

- Use Random Projection to a Subspace
  - similar to projection on the bottom k eigenvalues, but with different aim here
  - exploits redundancy in “Wide-Data” (high-dimensional data)
  - also powerful method in approximation algorithms (example: MAX-CUT)

- Why do we use a Random Projection?
  - If projection f is chosen deterministically, easy to find vectors u, v with $\|u - v\|$ large but $f(u) = f(v)$.

  $\Rightarrow$ Randomisation prevents the input to foil a specific deterministic algorithm

Generic Application of Preprocessing



Applications of the JL-Lemma

- Streaming Algorithms
- Preprocessing of many Machine Learning Methods

. . .

Random Projection, Margins, Kernels, and Feature-Selection

Avrim Blum

Department of Computer Science,
Carnegie Mellon University, Pittsburgh, PA 15213-3891

Abstract. Random projection is a simple technique that has had a number of applications in algorithm design. In the context of machine learning, it can provide insight into questions such as "why is a learning problem easier if data is separable by a large margin?" and "in what sense is choosing a kernel much like choosing a set of features?" This talk is intended to provide an introduction to random projection and to survey some simple learning algorithms and other applications to learning based on it. I will also discuss how, given a kernel as a black-box function, we can use various forms of random projection to extract an explicit small feature space that captures much of what the kernel is doing. This talk is based in large part on work in [BB05, BBV04] joint with Nina Balcan and Santosh Vempala.

Outline

Nearest Neighbour Algorithm

Dimensionality Reduction

Proof of JL-Lemma (advanced)

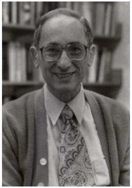
Chernoff Bounds and Concentration of Measure (Bonus Material)

Appendix

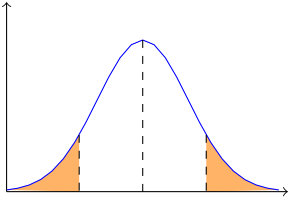
Chernoff Bounds

- Chernoffs bounds are "strong" bounds on the tail probabilities of sums of independent random variables (random variables can be discrete or continuous)
- usually these bounds decrease exponentially as opposed to a polynomial decrease in Markov's or Chebysheff's inequality (see example later)
- have found various applications in:
  - Random Projections
  - Approximation and Sampling Algorithms
  - Learning Theory (e.g., PAC-learning)
  - Statistics

  ⋮



Hermann Chernoff (1923-)



A Simple Chernoff Bound for Uniform Coin Flips

Uniform Chernoff Bound

Let $X_1, \ldots, X_n$ be independent random variables with $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[X_i = -1] = 1/2$. Let $X := \sum_{i=1}^{n} X_i$. Then for any $\lambda > 0$,

$$\mathbf{Pr}[X \geq \lambda] \leq e^{-\lambda^2/(2n)}.$$

A Simple Chernoff Bound for Uniform Coin Flips

Uniform Chernoff Bound

Let $X_1, \ldots, X_n$ be independent random variables with $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[X_i = -1] = 1/2$. Let $X := \sum_{i=1}^{n} X_i$. Then for any $\lambda > 0$,

$$\mathbf{Pr}[X \geq \lambda] \leq e^{-\lambda^2/(2n)}.$$

- This is a simple yet important setting, since r.v.'s are identical and symmetric
- Bound on $\mathbf{Pr}[X \leq -\lambda]$ follows by symmetry
- Bounds for the case $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[X_i = 0] = 1/2$ through substitution, see below

A Simple Chernoff Bound for Uniform Coin Flips

Uniform Chernoff Bound

Let $X_1, \ldots, X_n$ be independent random variables with $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[X_i = -1] = 1/2$. Let $X := \sum_{i=1}^n X_i$. Then for any $\lambda > 0$,

$$\mathbf{Pr}[X \geq \lambda] \leq e^{-\lambda^2/(2n)}.$$

- This is a simple yet important setting, since r.v.'s are identical and symmetric
- Bound on $\mathbf{Pr}[X \leq -\lambda]$ follows by symmetry
- Bounds for the case $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[X_i = 0] = 1/2$ through substitution, see below

Corollary

Let $X_1, \ldots, X_n$ be independent random variables with $\mathbf{Pr}[X_i = 0] = \mathbf{Pr}[X_i = 1] = 1/2$. Let $X := \sum_{i=1}^n X_i$ and $\mu := \mathbf{E}[X] = n/2$. Then for any $\lambda > 0$,

$$\mathbf{Pr}[X \geq \mu + \lambda] \leq e^{-2\lambda^2/n}.$$

Example: Repeated Uniform Coin Flips

Consider 100 independent coin flips. We wish to find an upper bound on the probability that the number of heads is greater or equal than 75.

Example: Repeated Uniform Coin Flips

Consider 100 independent coin flips. We wish to find an upper bound on the probability that the number of heads is greater or equal than 75.

- Markov's inequality: $X = \sum_{i=1}^{100} X_i$, $X_i \in \{0, 1\}$ and $\mathbf{E}[X] = 100 \cdot \frac{1}{2} = 50$.

$$\mathbf{Pr}[X \geq 3/2 \cdot \mathbf{E}[X]] \leq 2/3 = 0.666.$$

Example: Repeated Uniform Coin Flips

Consider 100 independent coin flips. We wish to find an upper bound on the probability that the number of heads is greater or equal than 75.

- Markov's inequality: $X = \sum_{i=1}^{100} X_i$, $X_i \in \{0, 1\}$ and $\mathbf{E}[X] = 100 \cdot \frac{1}{2} = 50$.

$$\mathbf{Pr}[X \geq 3/2 \cdot \mathbf{E}[X]] \leq 2/3 = 0.666.$$

- Chebyshev's inequality: $\mathbf{V}[X] = \sum_{i=1}^{100} \mathbf{V}[X_i] = 100 \cdot (1/2)^2 = 25$.

$$\mathbf{Pr}[|X - \mu| \geq t] \leq \frac{\mathbf{V}[X]}{t^2},$$

  and plugging in $t = 25$ gives an upper bound of $25/25^2 = 1/25 = 0.04$, much better than what we obtained by Markov's inequality.

Example: Repeated Uniform Coin Flips

Consider 100 independent coin flips. We wish to find an upper bound on the probability that the number of heads is greater or equal than 75.

- Markov's inequality: $X = \sum_{i=1}^{100} X_i$, $X_i \in \{0, 1\}$ and $\mathbf{E}[X] = 100 \cdot \frac{1}{2} = 50$.

$$\mathbf{Pr}[X \geq 3/2 \cdot \mathbf{E}[X]] \leq 2/3 = 0.666.$$

- Chebyshev's inequality: $\mathbf{V}[X] = \sum_{i=1}^{100} \mathbf{V}[X_i] = 100 \cdot (1/2)^2 = 25$.

$$\mathbf{Pr}[|X - \mu| \geq t] \leq \frac{\mathbf{V}[X]}{t^2},$$

  and plugging in $t = 25$ gives an upper bound of $25/25^2 = 1/25 = 0.04$, much better than what we obtained by Markov's inequality.

- The uniform Chernoff bound (Corollary) with $\mu = 50, \lambda = 25$ gives:

$$\mathbf{Pr}[X \geq \mu + \lambda] \leq e^{-2\lambda^2/100} = e^{-625/50} = e^{-12.5} = 0.00000372\ldots.$$

Example: Repeated Uniform Coin Flips

Consider 100 independent coin flips. We wish to find an upper bound on the probability that the number of heads is greater or equal than 75.

- Markov's inequality: $X = \sum_{i=1}^{100} X_i$, $X_i \in \{0,1\}$ and $\mathbf{E}[X] = 100 \cdot \frac{1}{2} = 50$.

$$\mathbf{Pr}[X \geq 3/2 \cdot \mathbf{E}[X]] \leq 2/3 = 0.666.$$

- Chebyshev's inequality: $\mathbf{V}[X] = \sum_{i=1}^{100} \mathbf{V}[X_i] = 100 \cdot (1/2)^2 = 25$.

$$\mathbf{Pr}[|X - \mu| \geq t] \leq \frac{\mathbf{V}[X]}{t^2},$$

  and plugging in $t = 25$ gives an upper bound of $25/25^2 = 1/25 = 0.04$, much better than what we obtained by Markov's inequality.

- The uniform Chernoff bound (Corollary) with $\mu = 50, \lambda = 25$ gives:

$$\mathbf{Pr}[X \geq \mu + \lambda] \leq e^{-2\lambda^2/100} = e^{-625/50} = e^{-12.5} = 0.00000372\ldots.$$

- the exact probability is $0.00000028\ldots$, so the Chernoff bound overestimates the actual probability by a factor of ≈ 10.

Example: Repeated Uniform Coin Flips

Consider 100 independent coin flips. We wish to find an upper bound on the probability that the number of heads is greater or equal than 75.

- Markov's inequality: $X = \sum_{i=1}^{100} X_i$, $X_i \in \{0, 1\}$ and $\mathbf{E}[X] = 100 \cdot \frac{1}{2} = 50$.

$$\mathbf{Pr}[X \geq 3/2 \cdot \mathbf{E}[X]] \leq 2/3 = 0.666.$$

- Chebyshev's inequality: $\mathbf{V}[X] = \sum_{i=1}^{100} \mathbf{V}[X_i] = 100 \cdot (1/2)^2 = 25$.

$$\mathbf{Pr}[|X - \mu| \geq t] \leq \frac{\mathbf{V}[X]}{t^2},$$

  and plugging in $t = 25$ gives an upper bound of $25/25^2 = 1/25 = 0.04$, much better than what we obtained by Markov's inequality.

- The uniform Chernoff bound (Corollary) with $\mu = 50, \lambda = 25$ gives:

$$\mathbf{Pr}[X \geq \mu + \lambda] \leq e^{-2\lambda^2/100} = e^{-625/50} = e^{-12.5} = 0.00000372\ldots.$$

- the exact probability is $0.00000028\ldots$, so the Chernoff bound overestimates the actual probability by a factor of ≈ 10.

Chernoff bound yields a more accurate result but needs independence!

Recipe for Deriving Chernoff Bounds

Main Steps

Basically, there are four main steps in deriving Chernoff bounds for sums of independent random variables $X = X_1 + \cdots + X_n$:

Recipe for Deriving Chernoff Bounds

Main Steps

Basically, there are four main steps in deriving Chernoff bounds for sums of independent random variables $X = X_1 + \cdots + X_n$:

1. Instead of working with X, we switch to e^{tX}, $t > 0$

Recipe for Deriving Chernoff Bounds

Main Steps

Basically, there are four main steps in deriving Chernoff bounds for sums of independent random variables $X = X_1 + \cdots + X_n$:

1. Instead of working with X, we switch to e^{tX}, $t > 0$
2. Apply Markov's inequality $\rightsquigarrow \mathbf{E}\left[\, e^{tX} \,\right]$

Recipe for Deriving Chernoff Bounds

Main Steps

Basically, there are four main steps in deriving Chernoff bounds for sums of independent random variables $X = X_1 + \cdots + X_n$:

1. Instead of working with X, we switch to e^{tX}, $t > 0$
2. Apply Markov's inequality $\rightsquigarrow$ $\mathbf{E}\left[\, e^{tX} \,\right]$
3. Compute/Estimate $\mathbf{E}\left[\, e^{tX} \,\right]$ using the independence of $X_1, \ldots, X_n$

Recipe for Deriving Chernoff Bounds

Main Steps

Basically, there are four main steps in deriving Chernoff bounds for sums of independent random variables $X = X_1 + \cdots + X_n$:

1. Instead of working with X, we switch to e^{tX}, $t > 0$
2. Apply Markov's inequality $\rightsquigarrow \mathbf{E}\left[\, e^{tX} \,\right]$
3. Compute/Estimate $\mathbf{E}\left[\, e^{tX} \,\right]$ using the independence of $X_1, \ldots, X_n$
4. Optimise the value of t to obtain best tail bound

Proof of the Uniform Chernoff Bound (1/2)

1. Switch to e^{tX}, $t > 0$:

Proof of the Uniform Chernoff Bound (1/2)

1. Switch to e^{tX}, $t > 0$:

$$\mathbf{Pr}\left[\, X \geq \lambda \,\right] = \mathbf{Pr}\left[\, e^{tX} \geq e^{t\lambda} \,\right]$$

Proof of the Uniform Chernoff Bound (1/2)

1. Switch to e^{tX}, $t > 0$:

$$\mathbf{Pr}\left[\,X \geq \lambda\,\right] = \mathbf{Pr}\left[\,e^{tX} \geq e^{t\lambda}\,\right]$$

2. Apply Markov's Inequality:

Proof of the Uniform Chernoff Bound (1/2)

1. Switch to e^{tX}, $t > 0$:

$$\mathbf{Pr}\left[X \geq \lambda \right] = \mathbf{Pr}\left[e^{tX} \geq e^{t\lambda} \right]$$

2. Apply Markov's Inequality:

$$\mathbf{Pr}\left[e^{tX} \geq e^{t\lambda} \right] \leq \frac{\mathbf{E}\left[e^{tX} \right]}{e^{t\lambda}}$$

Proof of the Uniform Chernoff Bound (1/2)

1. Switch to e^{tX}, $t > 0$:

$$\mathbf{Pr}\left[X \geq \lambda \right] = \mathbf{Pr}\left[e^{tX} \geq e^{t\lambda} \right]$$

2. Apply Markov's Inequality:

$$\mathbf{Pr}\left[e^{tX} \geq e^{t\lambda} \right] \leq \frac{\mathbf{E}\left[e^{tX} \right]}{e^{t\lambda}}$$

3. Estimate $\mathbf{E}\left[e^{tX} \right]$ using the independence of $X_1, \ldots, X_n$

Proof of the Uniform Chernoff Bound (1/2)

1. Switch to e^{tX}, $t > 0$:

$$\mathbf{Pr}\left[\, X \geq \lambda \,\right] = \mathbf{Pr}\left[\, e^{tX} \geq e^{t\lambda} \,\right]$$

2. Apply Markov's Inequality:

$$\mathbf{Pr}\left[\, e^{tX} \geq e^{t\lambda} \,\right] \leq \frac{\mathbf{E}\left[\, e^{tX} \,\right]}{e^{t\lambda}}$$

3. Estimate $\mathbf{E}\left[\, e^{tX} \,\right]$ using the independence of $X_1, \ldots, X_n$
   - First, we have

$$\mathbf{E}\left[\, e^{tX_i} \,\right] = \frac{1}{2}e^{t} + \frac{1}{2}e^{-t} \quad (= \cosh(t)).$$

Proof of the Uniform Chernoff Bound (1/2)

1. Switch to e^{tX}, $t > 0$:

$$\mathbf{Pr}\left[\, X \geq \lambda \,\right] = \mathbf{Pr}\left[\, e^{tX} \geq e^{t\lambda} \,\right]$$

2. Apply Markov's Inequality:

$$\mathbf{Pr}\left[\, e^{tX} \geq e^{t\lambda} \,\right] \leq \frac{\mathbf{E}\left[\, e^{tX} \,\right]}{e^{t\lambda}}$$

3. Estimate $\mathbf{E}\left[\, e^{tX} \,\right]$ using the independence of $X_1, \ldots, X_n$
   - First, we have

$$\mathbf{E}\left[\, e^{tX_i} \,\right] = \frac{1}{2} e^{t} + \frac{1}{2} e^{-t} \quad (= \cosh(t)).$$

   - Then

$$e^{t} = 1 + t + \frac{t^2}{2!} + \cdots \quad \text{and} \quad e^{-t} = 1 - t + \frac{t^2}{2!} - \cdots + (-1)^i \frac{t^i}{i!} + \cdots,$$

Proof of the Uniform Chernoff Bound (1/2)

1. Switch to e^{tX}, $t > 0$:

$$\mathbf{Pr}[X \geq \lambda] = \mathbf{Pr}\left[e^{tX} \geq e^{t\lambda}\right]$$

2. Apply Markov's Inequality:

$$\mathbf{Pr}\left[e^{tX} \geq e^{t\lambda}\right] \leq \frac{\mathbf{E}\left[e^{tX}\right]}{e^{t\lambda}}$$

3. Estimate $\mathbf{E}\left[e^{tX}\right]$ using the independence of $X_1, \ldots, X_n$
   - First, we have
   $$\mathbf{E}\left[e^{tX_i}\right] = \frac{1}{2}e^t + \frac{1}{2}e^{-t} \quad (= \cosh(t)).$$
   - Then
   $$e^t = 1 + t + \frac{t^2}{2!} + \cdots \quad \text{and} \quad e^{-t} = 1 - t + \frac{t^2}{2!} - \cdots + (-1)^i \frac{t^i}{i!} + \cdots,$$
   $$\Rightarrow \qquad \mathbf{E}\left[e^{tX_i}\right] = \sum_{i=0}^{\infty} \frac{t^{2i}}{(2i)!} \leq \sum_{i=0}^{\infty} \frac{(t^2/2)^i}{i!} = e^{t^2/2}.$$

Proof of the Uniform Chernoff Bound (1/2)

1. Switch to e^{tX}, $t > 0$:

$$\mathbf{Pr}[X \geq \lambda] = \mathbf{Pr}\left[e^{tX} \geq e^{t\lambda}\right]$$

2. Apply Markov's Inequality:

$$\mathbf{Pr}\left[e^{tX} \geq e^{t\lambda}\right] \leq \frac{\mathbf{E}\left[e^{tX}\right]}{e^{t\lambda}}$$

3. Estimate $\mathbf{E}\left[e^{tX}\right]$ using the independence of $X_1, \ldots, X_n$
   - First, we have

$$\mathbf{E}\left[e^{tX_i}\right] = \frac{1}{2}e^t + \frac{1}{2}e^{-t} \quad (= \cosh(t)).$$

   - Then

$$e^t = 1 + t + \frac{t^2}{2!} + \cdots \quad \text{and} \quad e^{-t} = 1 - t + \frac{t^2}{2!} - \cdots + (-1)^i \frac{t^i}{i!} + \cdots,$$

$$\Rightarrow \qquad \mathbf{E}\left[e^{tX_i}\right] = \sum_{i=0}^{\infty} \frac{t^{2i}}{(2i)!} \leq \sum_{i=0}^{\infty} \frac{(t^2/2)^i}{i!} = e^{t^2/2}.$$

since $(2i)! \geq 2^i \cdot i!$

Proof of the Uniform Chernoff Bound (1/2)

1. Switch to e^{tX}, $t > 0$:

$$\mathbf{Pr}[X \geq \lambda] = \mathbf{Pr}\left[e^{tX} \geq e^{t\lambda}\right]$$

2. Apply Markov's Inequality:

$$\mathbf{Pr}\left[e^{tX} \geq e^{t\lambda}\right] \leq \frac{\mathbf{E}\left[e^{tX}\right]}{e^{t\lambda}}$$

3. Estimate $\mathbf{E}\left[e^{tX}\right]$ using the independence of $X_1, \ldots, X_n$
   - First, we have

$$\mathbf{E}\left[e^{tX_i}\right] = \frac{1}{2}e^t + \frac{1}{2}e^{-t} \quad (= \cosh(t)).$$

   - Then

$$e^t = 1 + t + \frac{t^2}{2!} + \cdots \quad \text{and} \quad e^{-t} = 1 - t + \frac{t^2}{2!} - \cdots + (-1)^i \frac{t^i}{i!} + \cdots,$$

$$\Rightarrow \qquad \mathbf{E}\left[e^{tX_i}\right] = \sum_{i=0}^{\infty} \frac{t^{2i}}{(2i)!} \leq \sum_{i=0}^{\infty} \frac{(t^2/2)^i}{i!} = e^{t^2/2}.$$

since $(2i)! \geq 2^i \cdot i!$

   - Therefore,

$$\mathbf{E}\left[e^{tX}\right] = \prod_{i=1}^{n} \mathbf{E}\left[e^{tX_i}\right] = e^{t^2 n/2}.$$

Proof of the Uniform Chernoff Bound (2/2)

1.-3. The first three steps resulted in

$$\mathbf{Pr}\left[\, X \geq \lambda \,\right] = \mathbf{Pr}\left[\, e^{tX} \geq e^{\lambda} \,\right] \leq \frac{\mathbf{E}\left[\, e^{tX} \,\right]}{e^{t\lambda}} \leq e^{t^2 n/2 - t\lambda}.$$

Proof of the Uniform Chernoff Bound (2/2)

1.-3. The first three steps resulted in

$$\mathbf{Pr}\left[\, X \geq \lambda \,\right] = \mathbf{Pr}\left[\, e^{tX} \geq e^{\lambda} \,\right] \leq \frac{\mathbf{E}\left[\, e^{tX} \,\right]}{e^{t\lambda}} \leq e^{t^2 n/2 - t\lambda}.$$

4. To optimise t, let $f(t) = t^2 n/2 - t\lambda$, so

$$f'(t) = tn - \lambda,$$
$$f''(t) = t > 0.$$

Proof of the Uniform Chernoff Bound (2/2)

1.-3. The first three steps resulted in

$$\mathbf{Pr}[X \geq \lambda] = \mathbf{Pr}\left[e^{tX} \geq e^{\lambda}\right] \leq \frac{\mathbf{E}\left[e^{tX}\right]}{e^{t\lambda}} \leq e^{t^2 n/2 - t\lambda}.$$

4. To optimise t, let $f(t) = t^2 n/2 - t\lambda$, so

$$f'(t) = tn - \lambda,$$
$$f''(t) = t > 0.$$

Hence $t = \lambda/n$ is the minimum for $f(t)$

Proof of the Uniform Chernoff Bound (2/2)

1.-3. The first three steps resulted in

$$\mathbf{Pr}\left[\, X \geq \lambda \,\right] = \mathbf{Pr}\left[\, e^{tX} \geq e^{\lambda} \,\right] \leq \frac{\mathbf{E}\left[\, e^{tX} \,\right]}{e^{t\lambda}} \leq e^{t^2 n/2 - t\lambda}.$$

4. To optimise t, let $f(t) = t^2 n/2 - t\lambda$, so

$$f'(t) = tn - \lambda,$$
$$f''(t) = t > 0.$$

Hence $t = \lambda/n$ is the minimum for $f(t)$ and this yields

$$\mathbf{Pr}\left[\, X \geq \lambda \,\right] \leq e^{\lambda^2/(2n) - \lambda^2/n} = e^{-\lambda^2/(2n)}. \qquad \square$$

Extension to Non-Uniform Random Variables

Hoeffding's inequality

Let $X_1, \ldots, X_n$ be n independent random variables with $X_i \in [a_i, b_i]$ for each $1 \leq i \leq n$. Let $X := \sum_{i=1}^n X_i$ and $\mu = \mathbf{E}[X] = \sum_{i=1}^n \mathbf{E}[X_i]$. Then, for any $\delta \geq 0$,

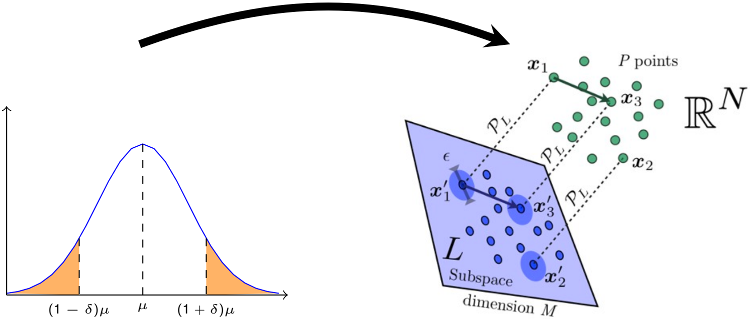
$$\mathbf{Pr}[|X - \mu| \geq \delta] \leq 2 \cdot \exp\left(-\frac{2\delta^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Further Extensions:

- Chernoff Bounds for sum of random variables with unbounded range (e.g., geometric random variables)
- Martingales (Azuma-Hoeffding's Inequality, Method of Bounded Independent Differences)
- Talagrand's Inequality (Measure Concentration and Expansion)
- ...

The last two extensions apply even to settings where the random variables are not independent.

Summary: Measure Concentration and Randomised Algorithms



- sums of independent random variables
- Chernoff Bounds: concrete tail inequalities that are exponential in the deviation
- Proof Method: Moment Generating Function & Markov's Inequality

- Random Projection Method
  - multiply by a random matrix
  - preserves distances up to $1 \pm \epsilon$
  - new dimension $\mathcal{O}(\log P/\epsilon^2)$

Outline

Appendix A: Moment Generating Functions

Moment-Generating Function

The moment-generating function of a random variable X is

$$M_X(t) = \mathbf{E}\left[e^{tX}\right], \qquad \text{where } t \in \mathbb{R}.$$

Appendix A: Moment Generating Functions

Moment-Generating Function

The moment-generating function of a random variable X is

$$M_X(t) = \mathbf{E}\left[e^{tX}\right], \qquad \text{where } t \in \mathbb{R}.$$

Using power series of e and differentiating shows that $M_X(t)$ encapsulates all moments of X.

Appendix A: Moment Generating Functions

Moment-Generating Function

The moment-generating function of a random variable X is

$$M_X(t) = \mathbf{E}\left[e^{tX}\right], \qquad \text{where } t \in \mathbb{R}.$$

Using power series of e and differentiating shows that $M_X(t)$ encapsulates all moments of X.

Lemma

1. If X and Y are two r.v.'s with $M_X(t) = M_Y(t)$ for all $t \in (-\delta, +\delta)$ for some $\delta > 0$, then the distributions X and Y are identical.
2. If X and Y are independent random variables, then

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

Appendix A: Moment Generating Functions

Moment-Generating Function

The moment-generating function of a random variable X is

$$M_X(t) = \mathbf{E}\left[e^{tX}\right], \qquad \text{where } t \in \mathbb{R}.$$

Using power series of e and differentiating shows that $M_X(t)$ encapsulates all moments of X.

Lemma

1. If X and Y are two r.v.'s with $M_X(t) = M_Y(t)$ for all $t \in (-\delta, +\delta)$ for some $\delta > 0$, then the distributions X and Y are identical.
2. If X and Y are independent random variables, then

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

Proof of 2:

$$M_{X+Y}(t) = \mathbf{E}\left[e^{t(X+Y)}\right] = \mathbf{E}\left[e^{tX} \cdot e^{tY}\right] \overset{(!)}{=} \mathbf{E}\left[e^{tX}\right] \cdot \mathbf{E}\left[e^{tY}\right] = M_X(t)M_Y(t) \qquad \square$$

References

- A. Blum, J. Hopcroft and R. Kannan
  Foundations of Data Science
  1st edition, Cambridge University Press, 2020.
  `https://www.cs.cornell.edu/jeh/book.pdf`
- J. Leskovec, A. Rajaraman, J. D. Ullmann
  Mining of Massive Datasets.
  3rd edition, Cambridge University Press, 2020.
- S. Shalev-Shwartz and S. Ben-David
  Understanding Machine Learning: From Theory to Algorithms
  Cambridge University Press, 2014.
  `https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf`