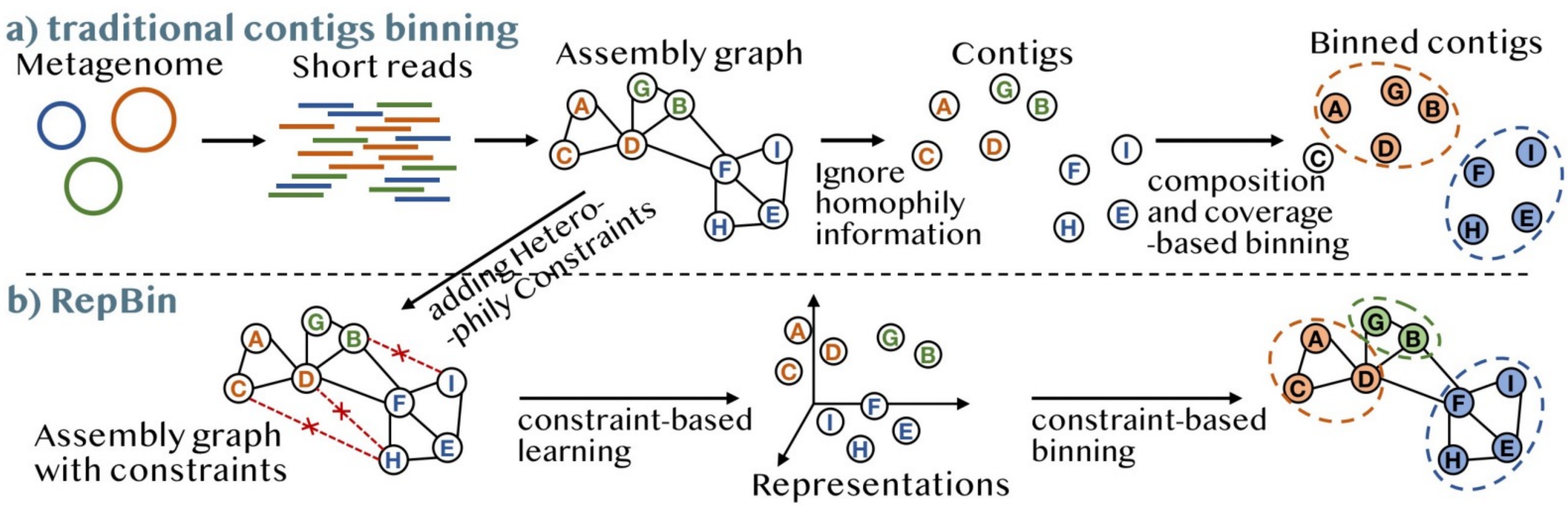
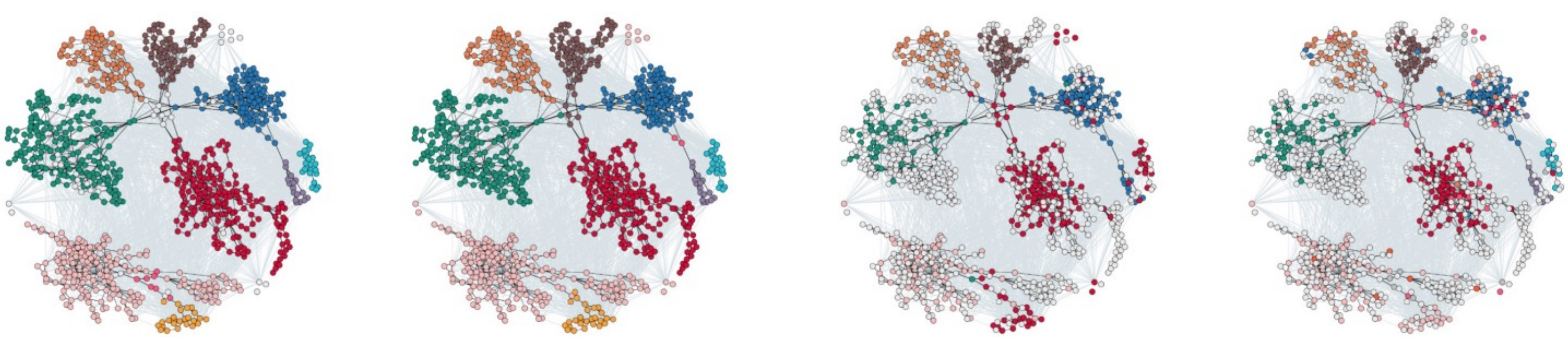
Review Lecture

Liang Zheng
Australian National University
liang.zheng@anu.edu.au

RepBin: Constraint-Based Graph Representation Learning for Metagenomic Binning Hansheng Xue, Vijini Mallawaarachchi, Yujia Zhang, Vaibhav Rajan, Yu Lin. AAAI 2022

- An unsupervised approach to discovering clusters, of genomic subsequences, associated with the unknown constituent organisms





GroundTruth[n=10] RepBin[n=10] BusyBeeWeb[n=6] CONCOCT[n=12]

Final exam

- **Time**: 9:30am – 13:00pm, Wednesday Nov 9th, 2021, AEDT.
- **Reading time**: 30 minutes.
- **Writing time**: 180 minutes
- **Zoom**: https://anu.zoom.us/j/85047332371?pwd=RDNOTnAvaG9VQnBDNzdGcnBlNWpWZz09

  Meeting ID: 850 4733 2371

  Password: 634639

- **Total marks**: 100 (it weights 60% of your mark in this course)

Final exam

Some other details will be published on Piazza.

- You can bring any material (paper-based, electronic)
- You can use a physical calculator or your computer calculator. You can do any programming to verify your answer.
- You can do any internet search, any typing, but no communication devices / software
- You must record your screen throughout the exam
- Phones are not allowed during the exam. (You can use it to take photos of your answer after the exam ends).
- The exam paper will be published on Wattle at 9:30am.

Final exam

- **Exam scope**:
- All the lectures (Week 1 to 12).
- Math and machine learning will roughly weight 40 and 60, respectively.
- **Exam questions**:
- Proofs, calculations, derivations and short answers.

Final exam

- **Exam questions**:
- For short answer questions
  - If asked, you need to provide a short explanation to your answer (e.g., yes or no).
  - If your explanation is incorrect, you will get 0.
- Write down everything on paper.
- No programming, no pseudo codes.
- Exam questions from past years are on Wattle.
- The difficulty will be on a similar level with past exams (may be slightly more difficult or easier, subject to students' perception).
- Percentage of easy, medium and hard questions: 20:60:20 (roughly)
- **It is not a hurdle**

Final exam

- Keep your camera on all the time, so that I can see your face.
- Keep muted.
  - Type your question in the chat **privately to me**.
  - You must not type publicly.
  - The only ones who can type publicly is the Lecturer and tutors.
- You can freely touch your keyboard
  - typing a question privately.
  - Searching the internet
  - No communication software/device!
- Prepare plenty of papers to write your answers on
  - If you like, you can use iPAD to write your answer, but you should make sure your iPAD is screen-recorded

Final exam

- Write different sections on separate papers
- On your paper, please clearly denote the question numbers, e.g., 1(a), 2(b). You don't need to specify the section number, as your answers will be submitted to the corresponding section portal.
- Stop writing at 13:00pm. You will then have 20 minutes to upload your answers.
  - Take pictures of your answers.
  - On wattle, you will see the submission portal for different exam sections.
  - You can submit either 1 file or multiple files to each portal.
  - If you submit 1 file, please name it as 1.pdf, which is the compilation of all your answers.
  - If you submit multiple files (maximum 20), you should name them as 1.pdf, 2.pdf, 3.pdf, …., according to the order of your answers.
  - You can also use other commonly seen files types, such as pdf, jpg, png.

Final exam

- Plagiarism in any form is absolutely forbidden.
- Probably fail the course if plagiarism is detected.
- There will be multiple versions of the exam questions
- According to our experience, because of the various question difficulties, some students couldn’t finish all the exam questions. It is normal if you find yourself in a similar situation. We will make adjustment accordingly.
- Please report to me through email if you find anything inappropriate.
- Type in the chat privately if you want to go to the restroom

Linear algebra

- Matrix operations
- Systems of linear equations (e.g., Gaussian elimination)
- Vector subspaces
- Linear combination and linear independence
- Basis of a vector space, rank
- Linear mappings and matrix

Matrix decomposition

- Understand determinant and its properties
- Being able to calculate determinant
- Eigenvalues, eigenvectors, eigenspaces, eigendecomposition and SVD
- How eigendecomposition simplifies matrix operations?

Matrix decomposition

- Determinant (defined for square matrices)
- For 2×2 matrices, if $\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, recall that the inverse of $\boldsymbol{A}$ is

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

- For $n = 3$ (known as Sarrus' rule),

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

- For a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ it holds that $\boldsymbol{A}$ is invertible if and only if $det(\boldsymbol{A}) \neq 0$.
- Properties of determinant
- We can use Gaussian elimination to compute $\det(\boldsymbol{A})$ by bringing $\boldsymbol{A}$ into row-echelon form. We can stop Gaussian elimination when we have $\boldsymbol{A}$ in a triangular form where the elements below the diagonal are all 0. Recall: the determinant of a triangular matrix is the product of the diagonal elements.

Eigenvalues and Eigenvectors

- Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{A}$ and $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ is the corresponding eigenvector of $\boldsymbol{A}$ if
$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$
- We call this equation the eigenvalue equation.
- We calculate eigenvalues as the root of the characteristic polynomial
$$p_{\boldsymbol{A}}(\lambda) := \det(\boldsymbol{A} - \lambda \boldsymbol{I})$$
- Eigenspace
- The eigenspace of $\boldsymbol{A}$ with respect to λ and is denoted by E_λ
- Eigendecomposition
- A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ can be factored into
$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$$

where $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{D}$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\boldsymbol{A}$, if and only if the eigenvectors of $\boldsymbol{A}$ form a basis of $\mathbb{R}^n$

Analytic geometry

- Understand norms, dot product and the more general inner product
- Lengths and distances are related to the inner product you choose
- Angles and orthogonality
- Linear mappings, eigenvalues, matrices, etc
- Orthogonal matrices $\boldsymbol{A}\,\boldsymbol{A}^T = \boldsymbol{I} = \boldsymbol{A}^T\boldsymbol{A}$ and rotations
- Orthonormal Basis
- Orthogonal projections and Gram-Schmidt Orthogonalization

Model meets data

- Understand model, predictor.
- Differentiate prediction, training and hyperparameter tuning.
- The role of cross validation
- Difference between parameters and hyperparameters
- Empirical risk minimization
- Evaluation metrics and loss functions
- The role of training data, testing data and validation data

Clustering

- Understanding EM algorithm for Clustering
- The difference between clustering and GMM
- Understanding the importance of key parameters in optimization.

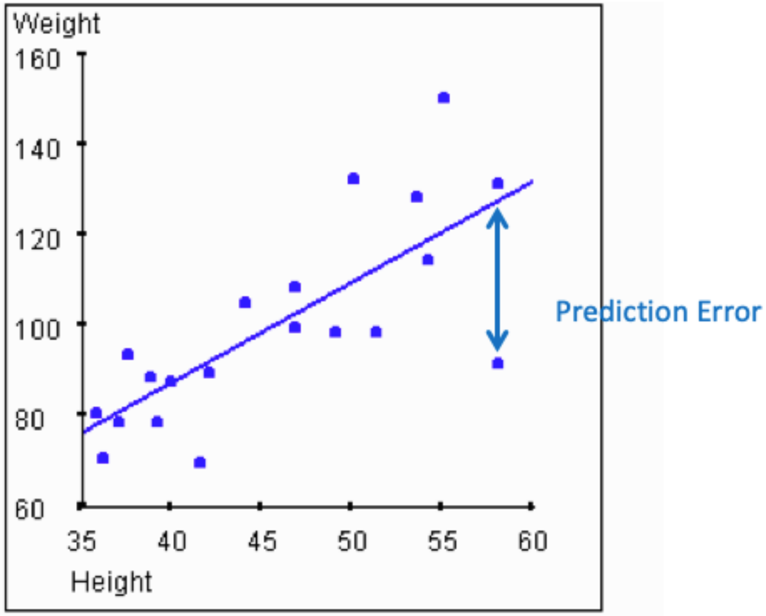
Vector calculus

- Being able to calculate the gradient / partial derivatives of a function (with respect to vectors or matrices) and apply them to optimize machine learning problems.
- Being able to use identities in lectures without proofs.

Linear regression

- What is linear regression?
- The relationship between linear regression, PCA and classification
- Being capable of deriving the gradient and the closed-form solution of linear regression
- What is linear regression with features? How can the features help linear regression?

Linear regression



- Model:

$$\boldsymbol{y}_n = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}_n$$

- Mean square error

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 = \frac{1}{N} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

- $\boldsymbol{X}$ is the data matrix; $\boldsymbol{y}$ is the label vector; $\boldsymbol{\theta}$ contains the parameters we want to optimize

Linear regression

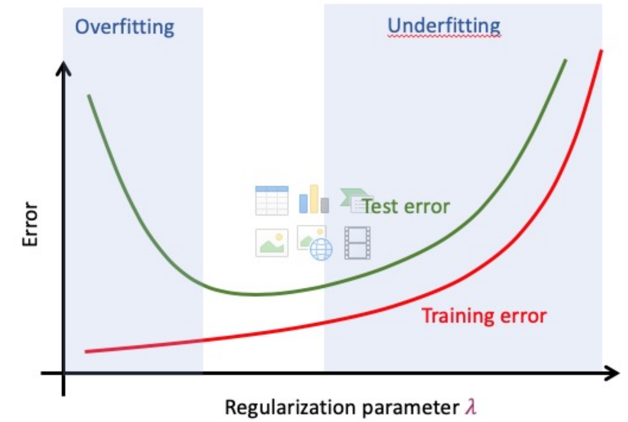
- We calculate the gradient of $\mathcal{L}$ with respect to the parameters $\boldsymbol{\theta}$ as

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\theta}} = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\left(\frac{1}{N}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\right)$$
$$= \frac{1}{N}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\theta}\right)$$
$$= \frac{1}{N}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\theta}\right)$$
$$= \frac{1}{N}\left(-2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} + 2\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) \in \mathbb{R}^{1\times D}$$

- The minimum is attained when the gradient is zero.

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\theta}} = \boldsymbol{0}^{\mathrm{T}} \;\Leftrightarrow\; \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}$$
$$\Leftrightarrow \quad \boldsymbol{\theta}^{\mathrm{T}} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$$
$$\Leftrightarrow \quad \boldsymbol{\theta} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Linear regression



- Regularized linear regression
- Loss function

$$\mathcal{L}_\lambda(\boldsymbol{\theta}) = \frac{1}{N}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$

- We calculate the gradient of $\mathcal{L}$ with respect to the parameters $\boldsymbol{\theta}$ as

$$\begin{aligned}\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\theta}} &= \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\left(\frac{1}{N}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|^2\right) \\ &= \frac{1}{N}\left(-2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} + 2\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) + \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}(\lambda\|\boldsymbol{\theta}\|^2) \\ &= \frac{1}{N}\left(-2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} + 2\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) + 2\lambda\boldsymbol{\theta}^{\mathrm{T}} \in \mathbb{R}^{1\times D}\end{aligned}$$

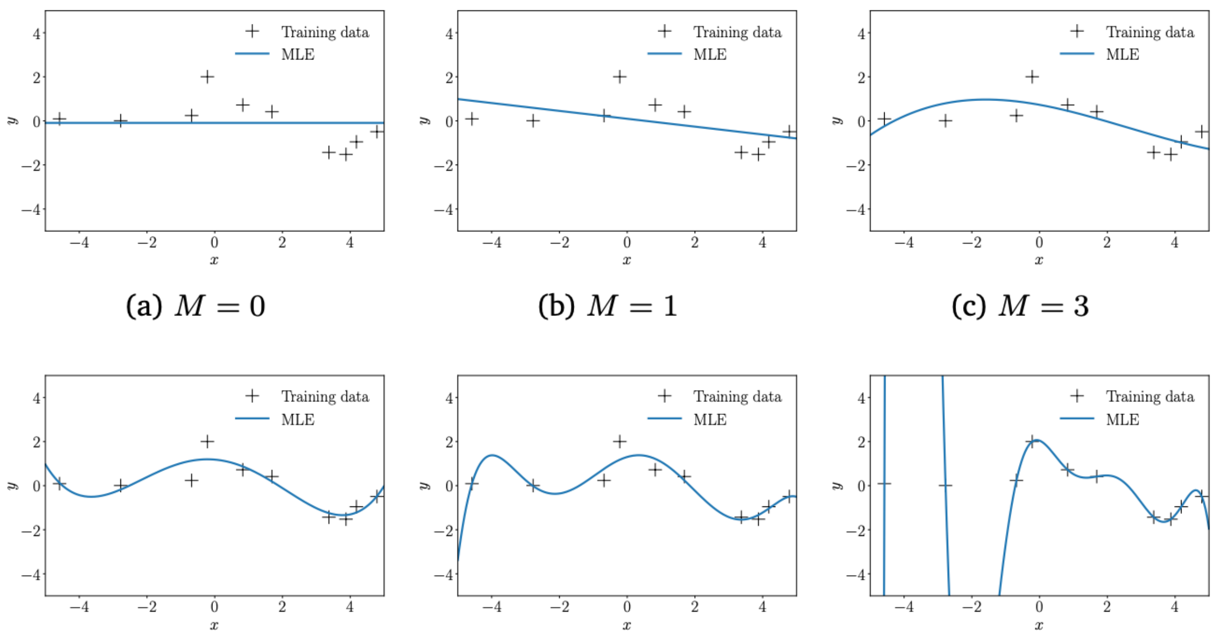
- The minimum is attained when the gradient is zero.

$$\begin{aligned}\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\theta}} = \boldsymbol{0}^{\mathrm{T}} &\Leftrightarrow\ 2\lambda\boldsymbol{\theta}^{\mathrm{T}} + \frac{1}{N}2\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \frac{1}{N}2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} \\ &\Leftrightarrow\ \boldsymbol{\theta}^{\mathrm{T}} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\left(N\lambda\boldsymbol{I} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \\ &\Leftrightarrow\ \boldsymbol{\theta} = \left(N\lambda\boldsymbol{I} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\end{aligned}$$

- Linear regression with features
- We are concerned with a regression problem $y = \phi^{\mathrm{T}}(x)\boldsymbol{\theta} + \epsilon$, where $x \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^K$. An example transformation that is used in this context is

$$\phi(x) = \begin{bmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \phi_{K-1}(x) \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \\ \vdots \\ x^{K-1} \end{bmatrix} \in \mathbb{R}^K$$

- We create a K-dimensional feature from a 1-dimensional input



(a) $M = 0$ (b) $M = 1$ (c) $M = 3$

(d) $M = 4$ (e) $M = 6$ (f) $M = 9$

Probability and distributions

- Being able to calculate discrete/continuous probability
- Being able to apply Bayes’ Theorem
- Familiar with Gaussian distributions
- Understand mean, variance, expectation, covariance…

Probability and distributions

- Discrete probabilities
- Sample question:
- I have a bag, containing 5 balls of the same size. There are 3 white balls and 2 red balls. I randomly pick a ball from my bag. After each pick, the ball is not put back to the bag. If I make two picks.
- Q1: what is the probability that the two picked balls are both white?
- Q2: what is the probability that at least one of the two picked balls are white?

Q1: $\frac{3}{5} \times \frac{2}{4} = \frac{3}{10}$

Q2: $1 - \frac{2}{5} \times \frac{1}{4} = \frac{9}{10}$

Sum Rule, Product Rule, and Bayes' Theorem

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \; \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$$

Sample question

In the population, 5% of the males are color blind, while 0.25% of the females are color blind. Now we randomly select a person and find he/she is color blind.

Q: what is the probability that this person is male?

A: We use random variable M to represent male, F to represent female, C to represent color blindness.

We have

$$P(C|M) = 0.05, P(C|F) = 0.0025.$$

Using Bayes' theorem,

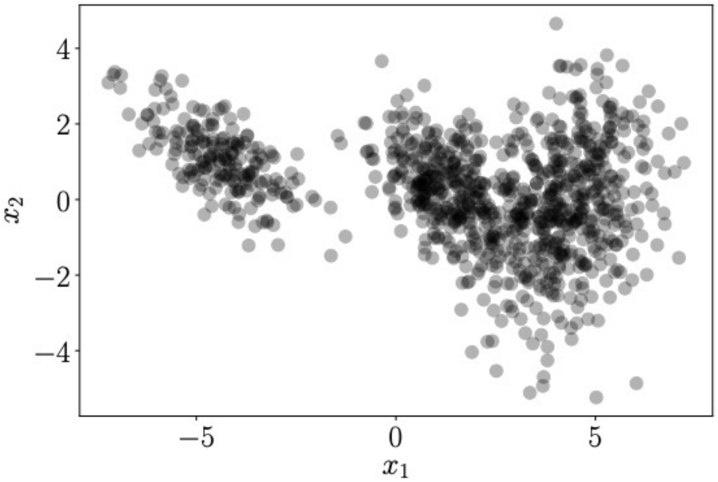
$$P(M|C) = \frac{P(M)P(C|M)}{P(M)P(C|M) + P(F)P(C|F)} = \frac{0.5\times0.05}{0.5\times0.05 + 0.5\times0.0025} = 95.2\%$$

Gaussian Mixture Models

- Understand GMMs, including
- Calculation with EM algorithm
- Understanding of the EM algorithm
- How to optimize the mean of the Gaussians

Gaussian Mixture Models

- A two-dimensional dataset



- A Gaussian mixture model (GMM) is a density model where we combine a finite number of K Gaussian distributions $N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ so that

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_\mathrm{k})$$

$$0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1$$

where we defined $\boldsymbol{\theta} := \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k \colon k = 1, \cdots, K\}$ as the collection of all parameters of the GMM.

EM Algorithm

- Initialize $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$. (below is an example)
  - $\pi_k = 1/K$ for all k
  - π_k: centroids from k-means algorithm
  - $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ the sample variance, for all k

- E-step: Evaluate responsibilities r_{nk} for every data point $\boldsymbol{x}_n$ using current parameters $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$:

$$r_{nk} = \frac{\pi_k \mathcal{N}\left(\boldsymbol{x}_n \middle| \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_j \pi_j \mathcal{N}\left(\boldsymbol{x}_n \middle| \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

- M-step: Re-estimate parameters $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ using the current responsibilities r_{nk} (from E-step):

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n$$

← You should be able to derive this

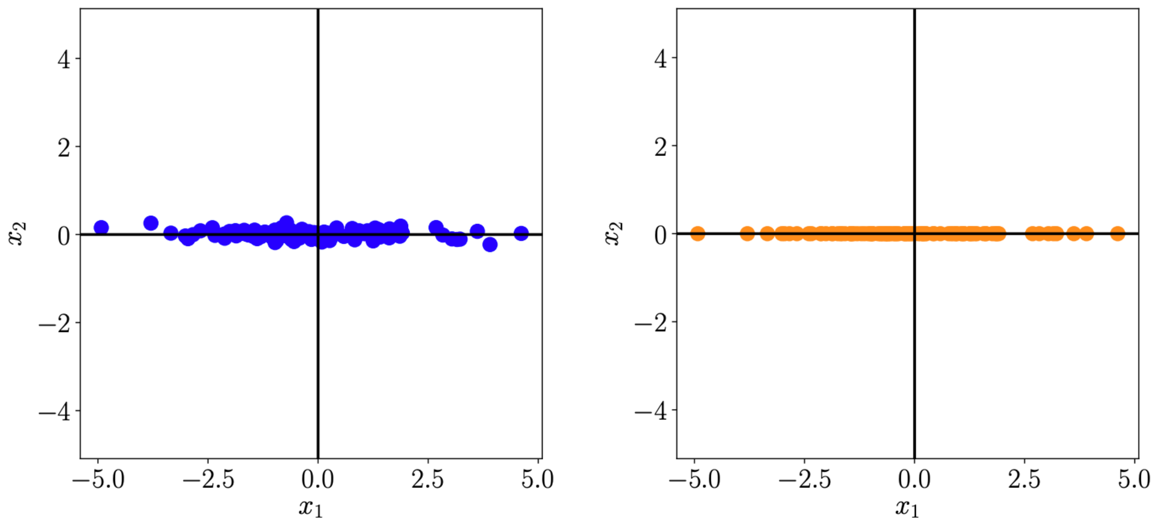
$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\pi_k = \frac{N_k}{N}$$

Principal Component Analysis

- PCA’s role in dimension reduction
- Understand the meaning of the intermediate data of PCA
- PCA calculation
- How PCA relates variance preservation with eigenvalues?
- PCA in high dimensions

Principal Component Analysis

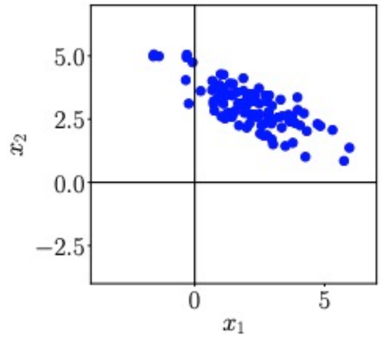


(a) Dataset with x_1 and x_2 coordinates.

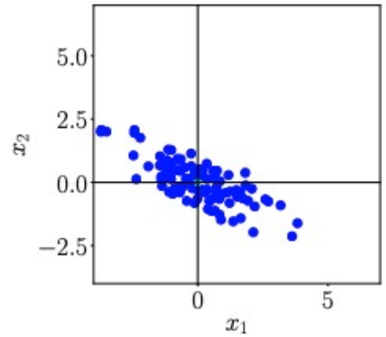
(b) Compressed dataset where only the x_1 coordinate is relevant.

PCA aims to find the direction where the variance is maximized.

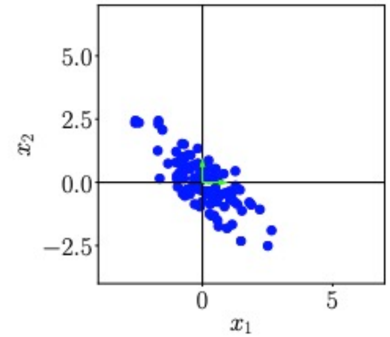
Key Steps of PCA in Practice



(a) Original dataset.

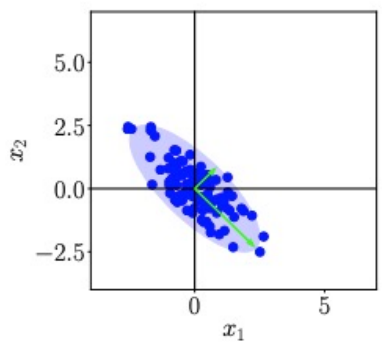


(b) Step 1: Centering by subtracting the mean from each data point.

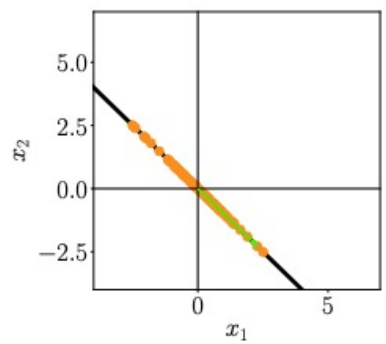


(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.

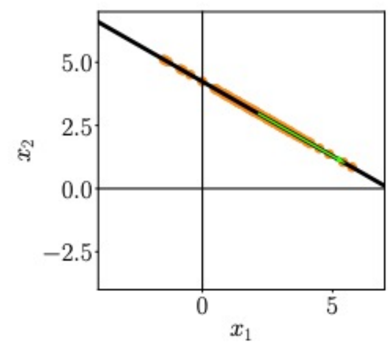
eigenface



(d) Step 3: Compute eigenvalues and eigenvectors (arrows) of the data covariance matrix (ellipse).



(e) Step 4: Project data onto the principal subspace.



(f) Undo the standardization and move projected data back into the original data space from (a).

eigendecomposition

PCA in high dimensions

- The covariance matrix could be very large
- Instead calculating

$$\boldsymbol{S} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \in \mathbb{R}^{D\times D}$$

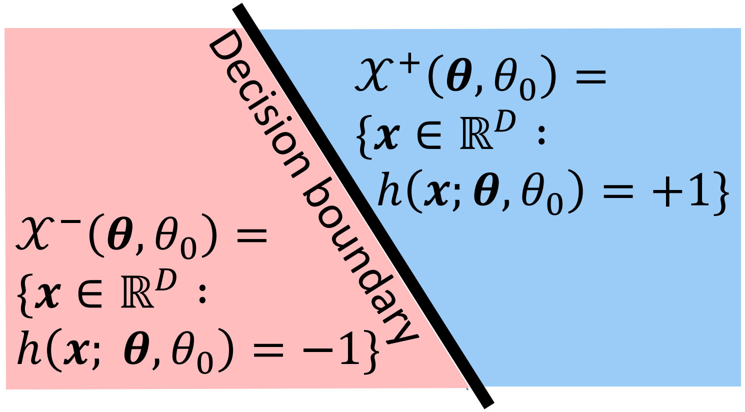
- We calculate

$$\widetilde{\boldsymbol{S}} = \frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$$

Classification

- Understand decision boundaries and decision regions
- Understand the differences between Perceptron, Hinge loss, and logistic loss
- Understand logistic regression
- Multi-way classification
- Differences and similarities between logistic regression and linear regression
- Logistic regression with regularization
- Being able to calculate gradient and apply gradient descent

Classification



linear classifier

non-linear classifier

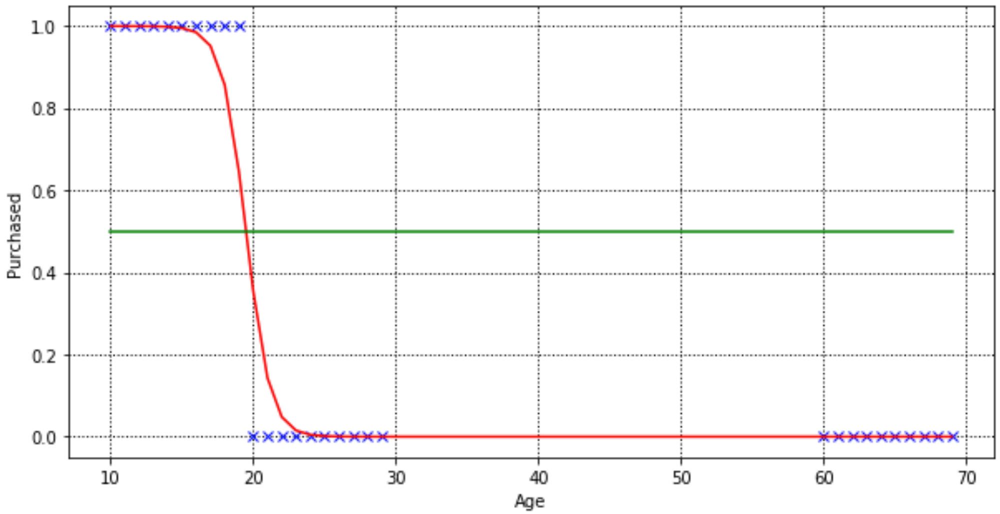
A classifier h partitions the space into decision regions that are separated by decision boundaries. In each region, all the points map to the same label. Many regions could have the same label.

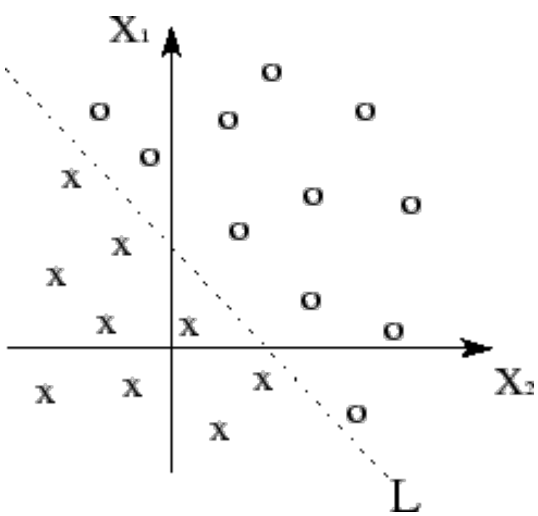
Classification

- The training data $\mathcal{D}$ is linearly ∈ separable if there exist parameters $\boldsymbol{\theta} = [\theta_0, \theta_1, \theta_2, \dots, \theta_D]^T$ such that for all $(\boldsymbol{x}, y) \in \mathcal{D}$,

$$y(\boldsymbol{\theta}^\top x) > 0$$

Logistic regression vs linear regression





Perceptron

Rosenblatt (1962)

Classifier:

$$h(\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{sign}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}\right)$$

Let $\mathcal{L}_1(\boldsymbol{\theta};\boldsymbol{x},y) = 1$ (0 otherwise) if

- $y \neq h(\boldsymbol{x};\boldsymbol{\theta})$, or [misclassified]
- (x,y) is on decision boundary [boundary]

Note that $y\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}\right) \leq 0$ if

- $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}$ and y differ in sign, or [misclassified]
- $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}$ is zero [boundary]

$$\mathcal{L}_1(\boldsymbol{\theta};\boldsymbol{x},y) = \left[\!\left[y\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}\right) \leq 0 \right]\!\right] = \mathrm{Loss}\left(y\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}\right)\right)$$

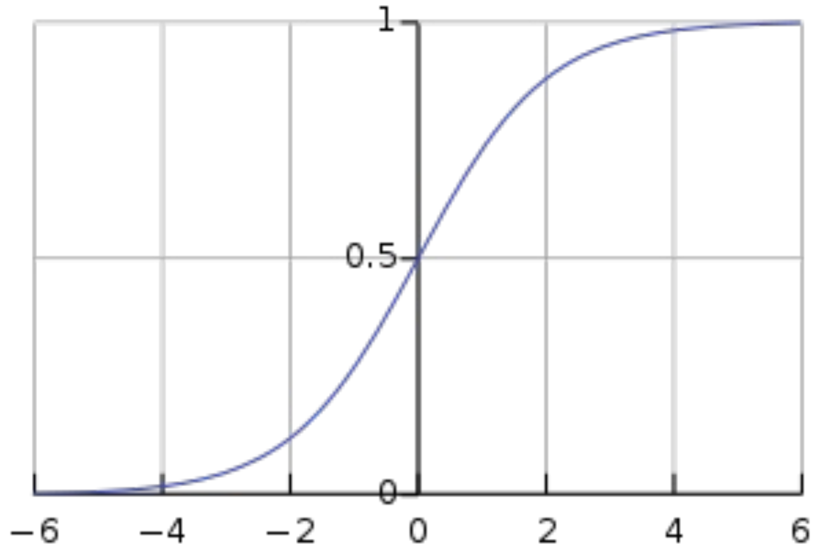
where $\mathrm{Loss}(z) = [\![z \leq 0]\!]$ is the zero-one loss.

Logistic regression

- $\mathbb{P}(\, y = +1 \mid \boldsymbol{x} \,) = \mathrm{sigmoid}(\boldsymbol{\theta}^\top \boldsymbol{x}) = \mathrm{sigmoid}\big(\mathrm{y}(\boldsymbol{\theta}^\top \boldsymbol{x})\big)$
- $\mathrm{sigmoid}(t) = \frac{1}{1+e^{-t}}$

$$h(\boldsymbol{x}; \boldsymbol{\theta}) \geq \frac{1}{2} \iff \mathrm{sigmoid}(\boldsymbol{\theta}^\top \boldsymbol{x}) \geq \frac{1}{2} \iff \boldsymbol{\theta}^\top \boldsymbol{x} \geq 0$$

$$h(\boldsymbol{x}; \boldsymbol{\theta}) < \frac{1}{2} \iff \mathrm{sigmoid}(\boldsymbol{\theta}^\top \boldsymbol{x}) < \frac{1}{2} \iff \boldsymbol{\theta}^\top \boldsymbol{x} < 0$$



Logistic Loss

Minimize the training loss

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= -\frac{1}{N} \log L(\boldsymbol{\theta}; \mathcal{D}) \\
&= -\frac{1}{N} \log \prod_{(\boldsymbol{x},y)\in\mathcal{D}} \mathbb{P}(y|\boldsymbol{x}) \\
&= -\frac{1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log \mathbb{P}(y|\boldsymbol{x}) \\
&= -\frac{1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log \frac{1}{1+e^{-y(\boldsymbol{\theta}^\top \boldsymbol{x})}} \\
&= \frac{1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log\left(1 + e^{-y(\boldsymbol{\theta}^\top \boldsymbol{x})}\right) \\
&= \frac{1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \operatorname{Loss}\left(y(\boldsymbol{\theta}^\top \boldsymbol{x})\right)
\end{aligned}
$$

$\operatorname{Loss}(z) = \log(1 + e^{-z})$ is the *logistic loss*.

Regularized Logistic Regression

- When your data has a high-dimensional feature, or your training set is small, you might have the over-fitting problem.

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log\left(1 + e^{-y(\boldsymbol{\theta}^\top \boldsymbol{x})}\right) + \frac{\lambda}{2N} \sum_{j=1}^{D} \theta_j^2$$

- We are doing regularization on $\theta_1, \theta_2, \dots, \theta_D$
- When using gradient descent, we have

$$\theta_0 \longleftarrow \theta_0 - \frac{\eta_k}{N} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} x^{(0)} (h(\boldsymbol{x};\boldsymbol{\theta}) - [\![y = 1]\!])$$

$$\theta_j \longleftarrow \theta_j - \frac{\eta_k}{N} \left[\sum_{(\boldsymbol{x},y)\in\mathcal{D}} x^{(j)} (h(\boldsymbol{x};\boldsymbol{\theta}) - [\![y = 1]\!]) + \lambda\, \theta_j \right]$$

$$j = 1,2, \dots, D$$