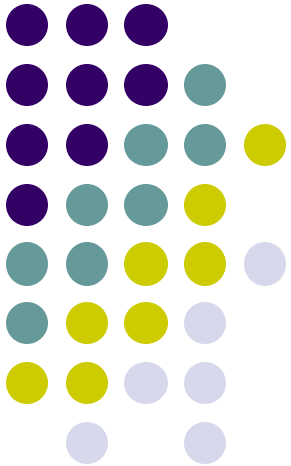
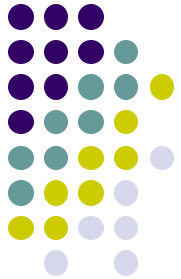
ErSE222: Machine learning in Geoscience

Feb. 2nd, 2025

Tariq Alkhalifah and Omar Saad

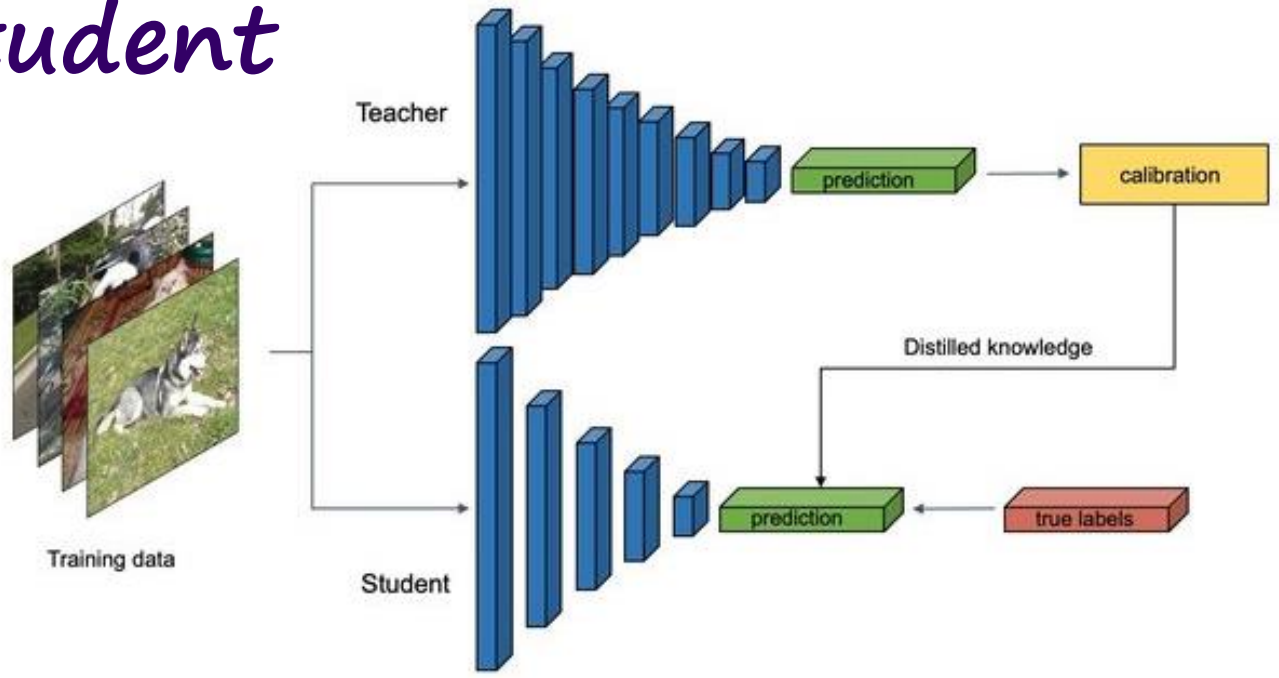


Important terms in ML



- Model or hypothesis
- Model parameters, weights and biases
- Model architecture
- Training: pretraining, fine tuning and Pruning (optimization)
- Training data, validation, testing, labeled , labelless
- Type of training: supervised/unsupervised/self-supervised
- Overfitting/generalization
- Data bias
- Inductive bias
- Pruning, bagging, boosting, …..
- Regularization
- Bayesian NN
- Teacher, student, teacher forcing!

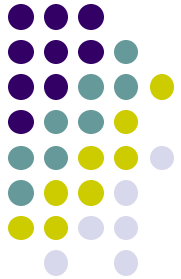
Teacher-Student models



Recent years have witnessed dramatically improvements in the knowledge distillation, which can generate a compact student model for better efficiency while retaining the model effectiveness of the teacher model. Previous studies find that: more accurate teachers do not necessary make for better teachers due to the mismatch of abilities. In this paper, we aim to analysis the phenomenon from the perspective of model calibration. We found that the larger teacher model may be too over-confident, thus the student model cannot effectively imitate. While, after the simple model calibration of the teacher model, the size of the teacher model has a positive correlation with the performance of the student model.

Yang L. and Song J., 2021, Rethinking the Knowledge Distillation From the Perspective of Model) Calibration

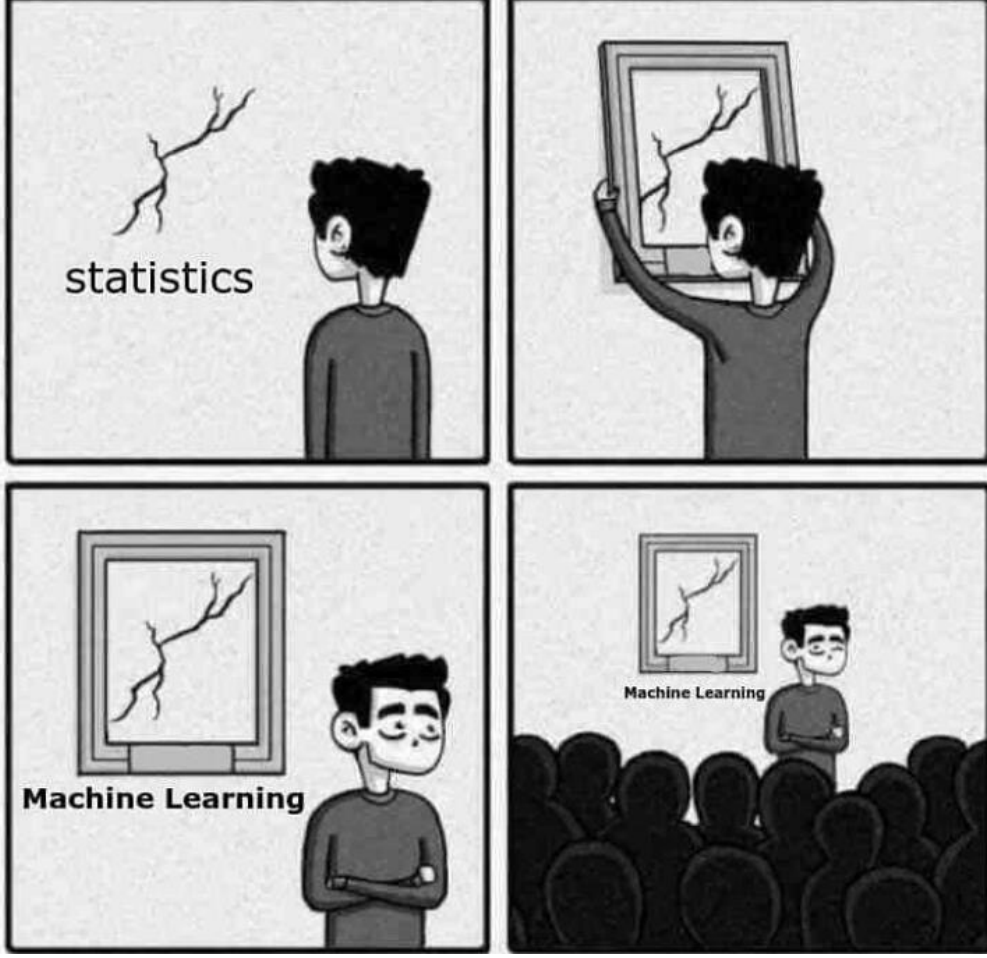
Beyond the hype, what is ML?

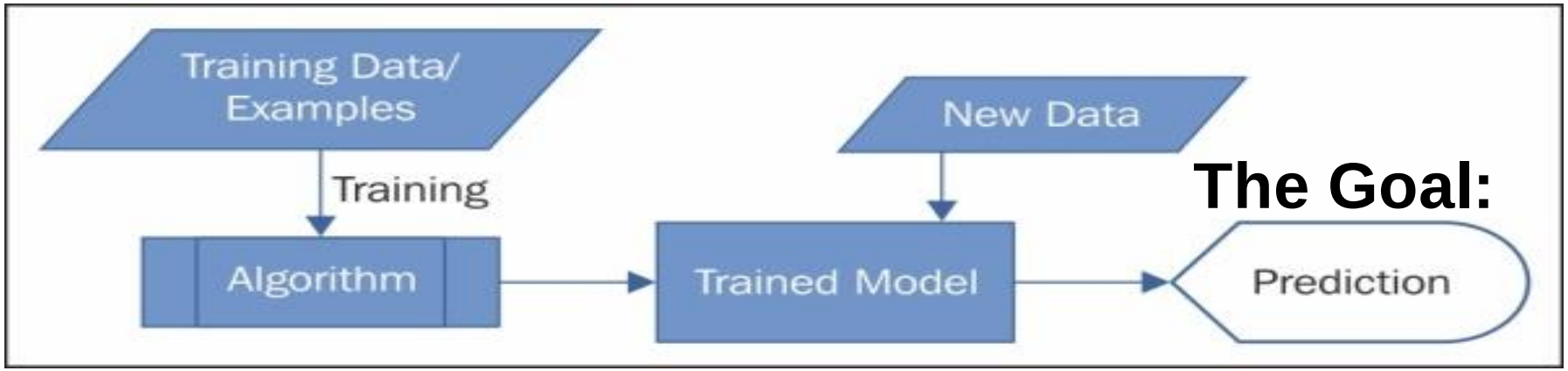


- It is a numerical tool, like any other tool, but it a slightly more extensive form based on optimization.
- Example: we do not hand craft the transformations, we learn them.
- Data driven, but possibly physics.
- Interpolation and extrapolation
- Based on learned features (optimal basis): MLP, CNN, Attention,….

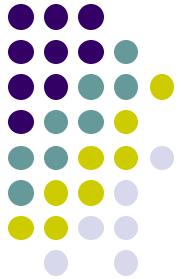
The ingredients

- *Linear Algebra*
- *Probability and Statistics*
- *Calculas*
- *Optimization*
- *Numerical methods*
- *Computer science*





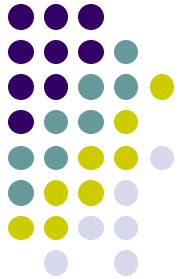
Why Linear Algebra?



- Linear algebra is the study of vectors, and matrices, as their transforms or operators (acting like functions).

- In machine learning, it is used to:
  - Represent data and models.
  - Perform operations like scaling, rotations, and projections.
  - Enable efficient computation for algorithms.

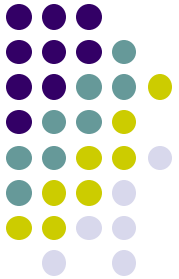
Vectors



- Represent data points: Each feature as a dimension (e.g., [Age, Salary, Experience]).
- Parameters in models: Weights in linear regression or deep learning.
- Key operations: Dot product (similarity), vector norms (magnitude), element-wise multiplications.
- $v \in R^n$, spans dimension n

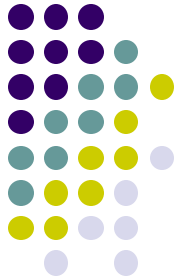
  L2 norm: $||\mathbf{v}|| = \sqrt{(v_1^2 + v_2^2 + ... + v_n^2)}$.

Matrices



- Matrices organize data and transformations:
  - Rows as samples, columns as features.
  - Transformations such as scaling and rotations.
  - Represent relationships: **XW** (inputs × weights).
- Example: A data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, where m = samples (vectors), n = features.
- Transformations changes the vectors direction and magitude.

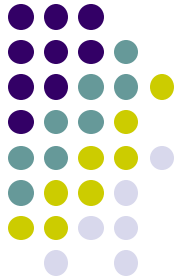
Matrix Operations



- Matrix operations are critical for ML models:
  - Addition and subtraction: Combine data sources.
  - Multiplication: Transformations and layer computations.
  - Transpose: Reorganizing data ($\mathbf{X}^T$).

- Example: $\mathbf{y} = \mathbf{XW} + \mathbf{b}$ (linear model prediction).

Matrix Inverse and Rank



- Matrix inverse ($\mathbf{A}^{-1}$) and rank play vital roles:
  - Inverse: Solves $\mathbf{Ax} = \mathbf{b}$ (e.g., regression normal equation).
  - Rank: Measures feature independence and dataset quality.

- Full rank ensures no redundant features.

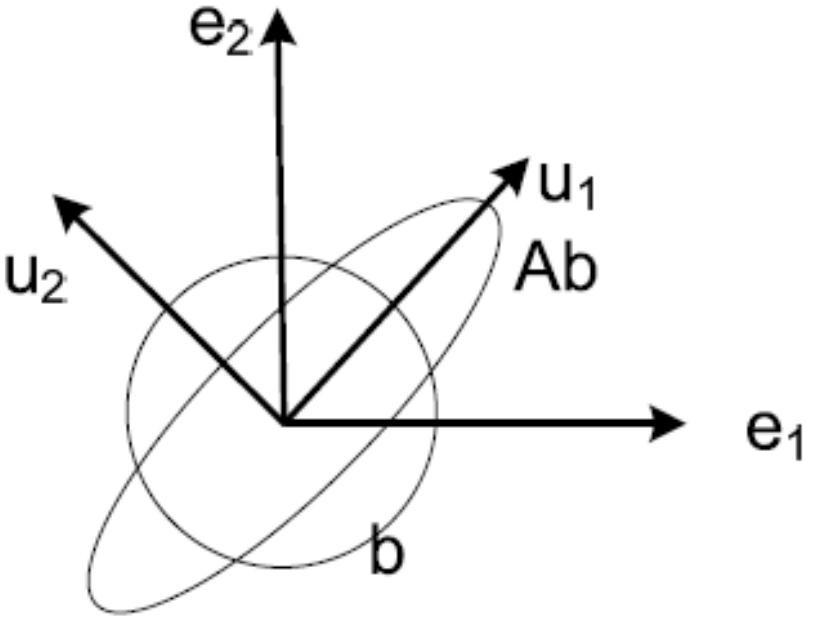
Eigenvalues and eigenvector

- For a special vector:

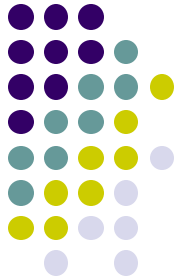
  $\mathbf{Au}=\lambda\mathbf{u}$

- λ: eigenvalue
- **u**: eigenvector

In $\mathbb{R}^2$

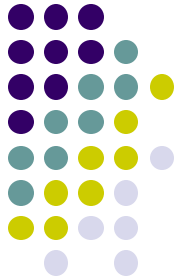


Eigenvalues and Eigenvectors



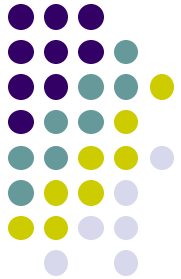
- Eigenvalues and eigenvectors describe transformations:
  - Eigenvector: Direction unchanged by the transformation.
  - Eigenvalue: Scale factor for the eigenvector.
- Applications:
  - PCA: Find dominant components.
  - Spectral clustering and graph ML.

Singular Value Decomposition (SVD)



- SVD decomposes a matrix **A** as: $\mathbf{A} = \mathbf{U\Sigma V}^T$
  - **U**: Orthogonal matrix of left singular vectors.
  - **Σ**: Diagonal matrix of singular values.
  - $\mathbf{V}^T$: Orthogonal matrix of right singular vectors.
- Applications: PCA, noise reduction, latent features.

Principal Component Analysis (PCA)



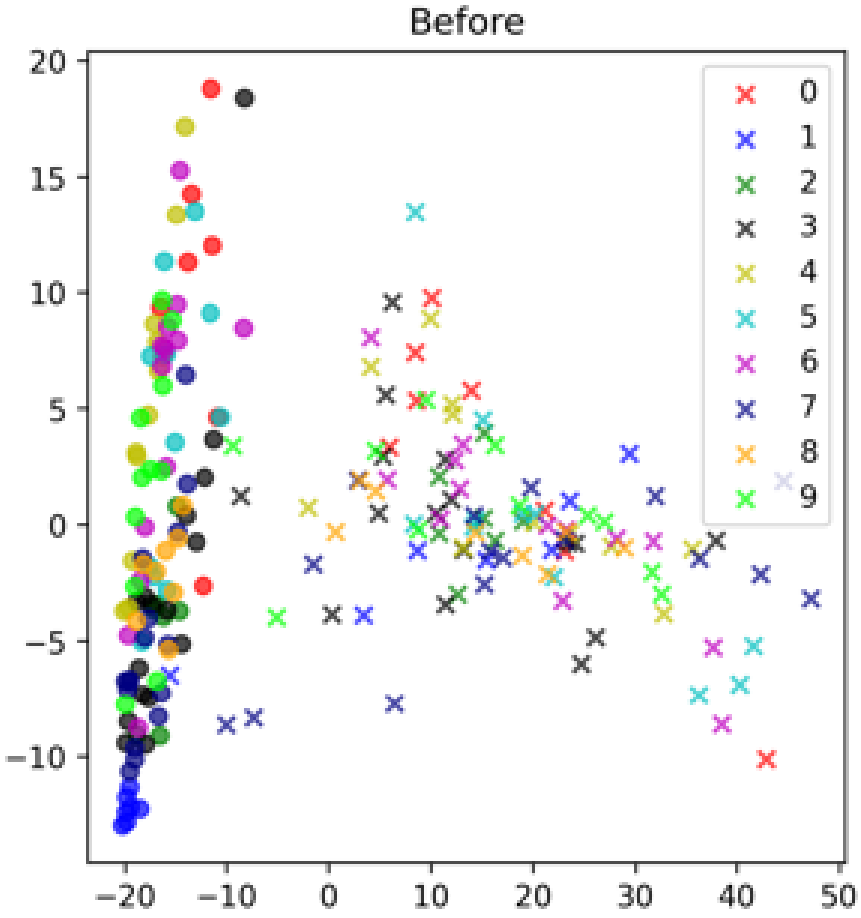
- PCA uses linear algebra to reduce dimensions.

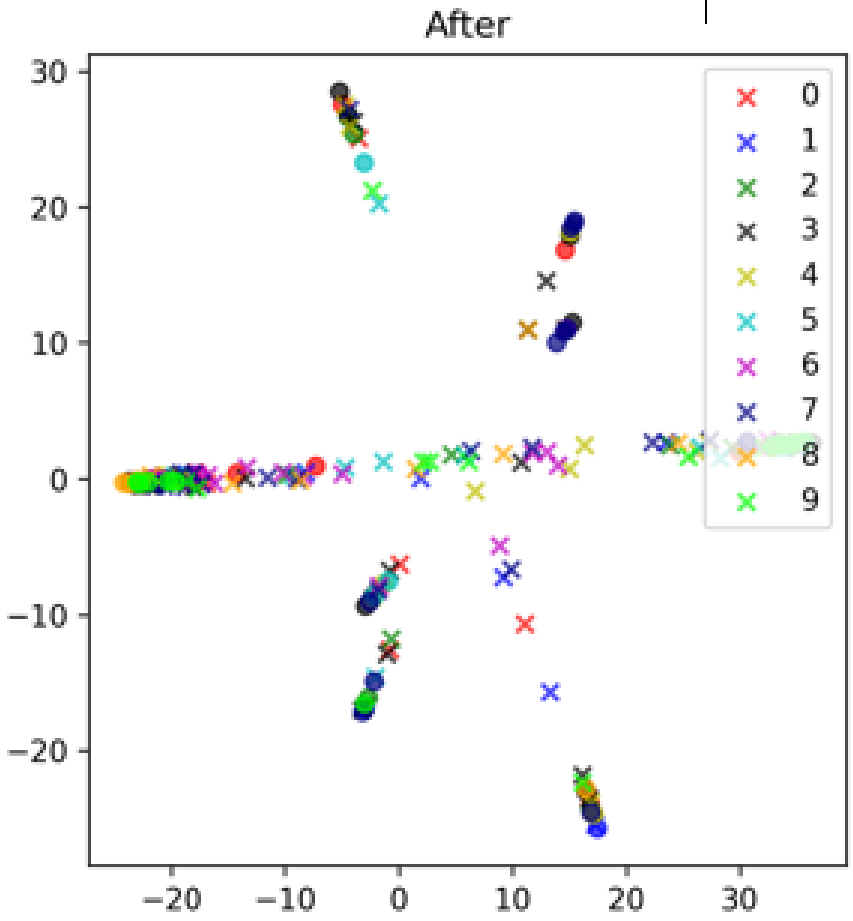
Steps:

1. Center data by subtracting the mean.
2. Compute covariance matrix.
3. Perform eigenvalue decomposition.
4. Select top-k eigenvectors (principal components).

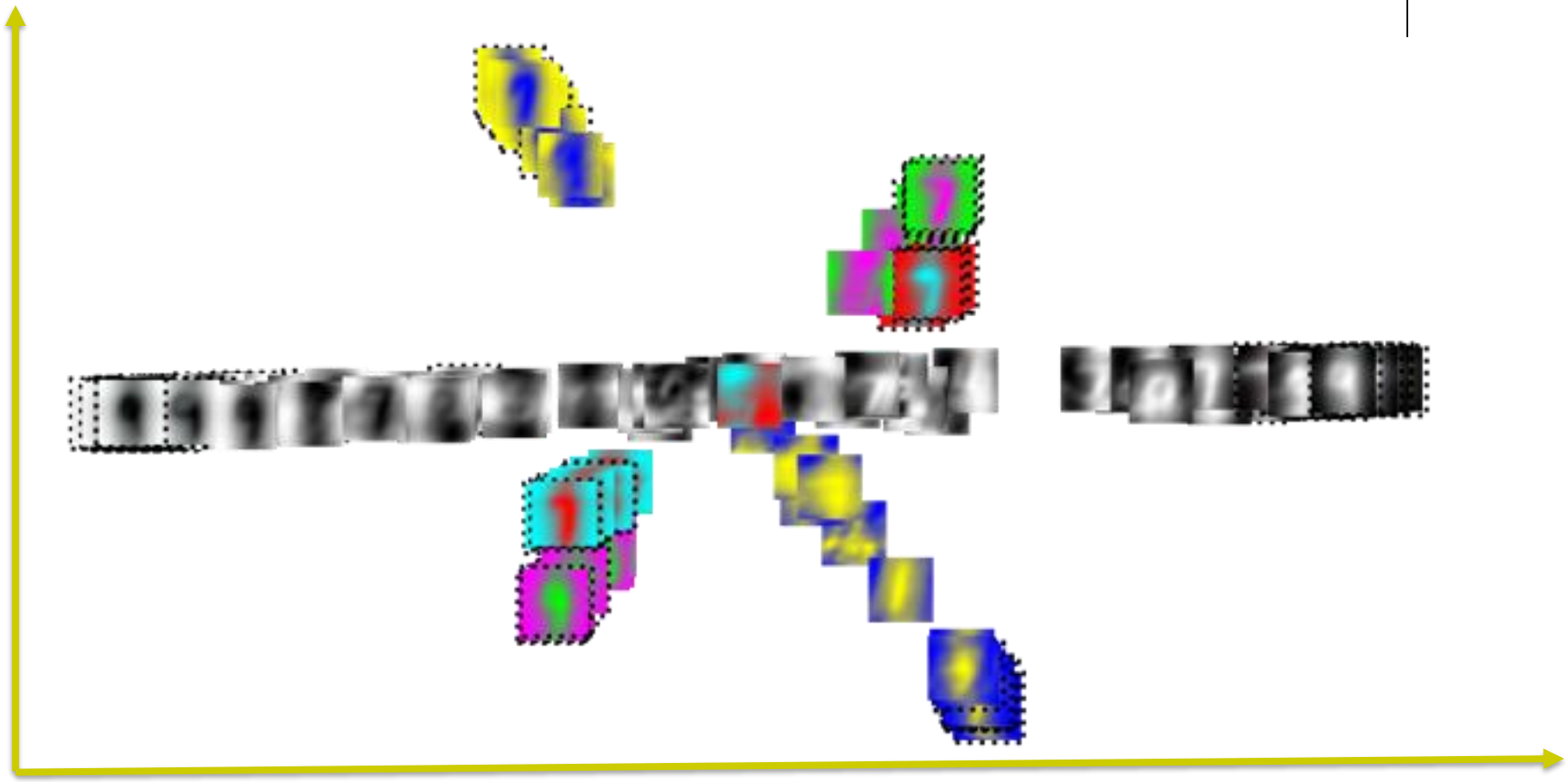
PCA with images (before)

Principal component analysis

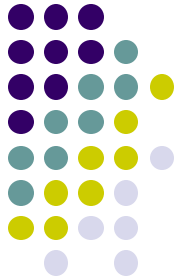




PCA with images (after)

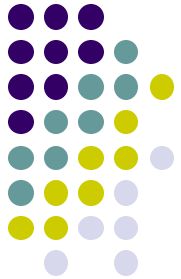


Norms and Regularization



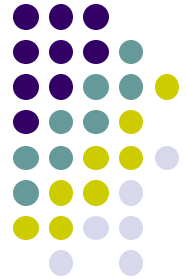
- Norms measure vector or matrix size:
  - L1 norm: $||\mathbf{x}||_1 = \sum|x_i|$, promotes sparsity (Lasso Regression).
  - L2 norm: $||\mathbf{x}||_2 = \sqrt{\sum x_i^2}$, penalizes large weights (Ridge Regression).
- Regularization reduces overfitting by constraining model parameters.

Pseudoinverse



- Pseudoinverse A^+ solves systems where A is not invertible:
  - Least-squares solution: $\mathbf{x} = (\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T\mathbf{y}$.
  - Handles overdetermined and underdetermined systems.
- Widely used in linear regression.

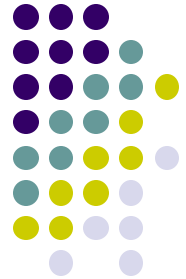
Feature Independence and Rank



- Full-rank matrices ensure:
  - Independent features (no redundancy).
  - Robust model training.

- Check: For $\mathbf{X} \in \mathbb{R}^{m \times n}$, Rank($\mathbf{X}$) = min(m, n).

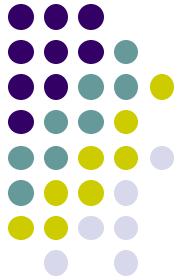
Linear Regression as Linear Algebra



- Linear regression is solved using $\mathbf{Ax} = \mathbf{b}$:
  - $\mathbf{X}$: Feature matrix $\mathbf{A}$.
  - θ: Weight vector $\mathbf{x}$.
  - $\mathbf{y}$: Target vector $\mathbf{b}$.

- Normal equation: $\theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Neural Networks

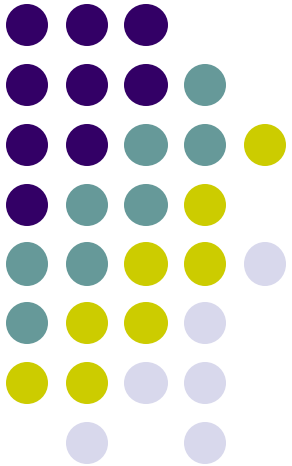


- Neural networks rely on:
  - Weight matrices: **W** connects layers.
  - Bias vectors: Added to layer outputs.

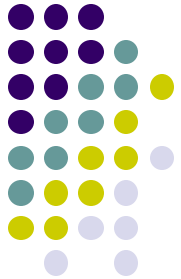
- Forward pass: **Z** = **XW** + **b** → Apply activation (e.g., ReLU, sigmoid).

Introduction to Probability & Statistics for Machine Learning in Geoscience

A foundational overview with key concepts, equations, and examples



Covering

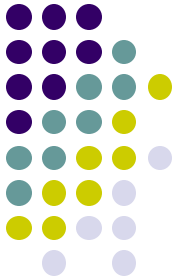


- Samples & Probability Theory
- Distributions
- Joint, Marginal & Conditional Distributions
- Bayesian Theory & Inference
- Sampling from Posterior

Definitions

- Experiment: A repeatable process that produces a well-defined outcome. Like, recording seismic activity, data or Earthquake magnitude.
- Sample Space (Ω): The set of all possible outcomes of an experiment, e.g., all possible earthquake magnitudes. It could be finite or infinite, and it could be countable or uncountable.
- Event (E): A subset of the sample space that represents an outcome or set of outcomes of interest, like a single measurement.

Probability Theory



$$P(E) = \frac{\text{Number of } E \text{ outcomes}}{\text{Total number of outcomes}}$$

$$\int_{-\infty}^{\infty} p(y)dy = 1, p(y) \geq 0$$

Examples: Normal, Uniform, Exponential, Poisson, etc.

Random Variables

- A random variable is a function that assigns a numerical value to each outcome of an experiment.
- Random variables are classified as ”discrete” or ”continuous”.
- For discrete: Probability Mass Function (PMF): *p(x) = P(X = x).*
- For continuous: $f(y)$ represents the density (PDF) of Y around y, then

$$P(a \leq Y \leq b) = \int_a^b f(y)dy$$

The expectation

- The expectation (or expected value) of a random variable X, denoted as E[X], represents the mean or average value that X takes.

- For a discrete random variable:

$$E[X] = \Sigma \; x \; P(X = x)$$

- For a continuous random variable:

$$E[X] = \int x \, f(x) \, dx$$

The variance

- Variance measures the spread of a random variable X around its expectation E[X].
- For a discrete random variable:

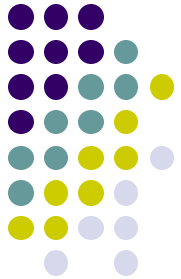
$$\mathrm{Var}(X) = E[(X - E[X])^2] = \Sigma\, (x - E[X])^2\, P(X = x)$$

- For a continuous random variable:

$$\mathrm{Var}(X) = \int (x - E[X])^2\, f(x)\, dx$$

- The square root of variance is called the standard deviation: $\sigma(X) = \sqrt{\mathrm{Var}(X)}$.

The Joint distribution



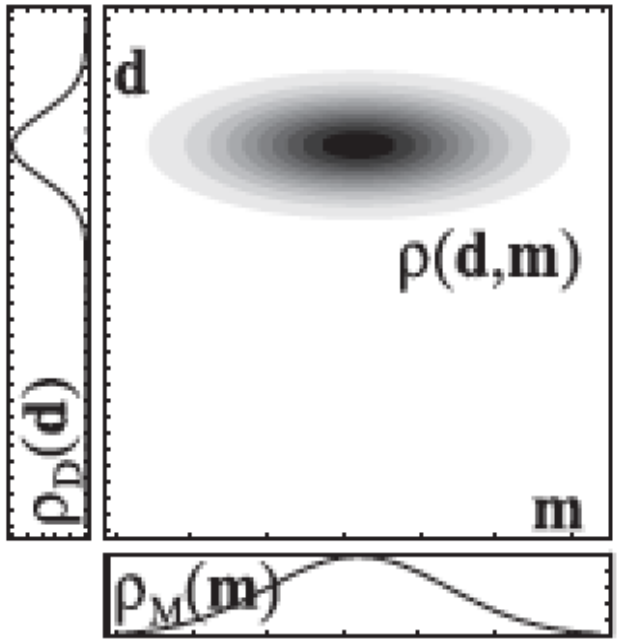
- Describes the probability of two or more random variables taking specific values simultaneously.
- The probability density function (PDF) is given $f_{X,Y}(x, y)$

The marginal distribution

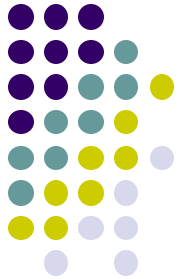
- of a random variable is obtained by summing (or integrating) the joint distribution over the other variable(s).
- So the PDF:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$$



The conditional distribution

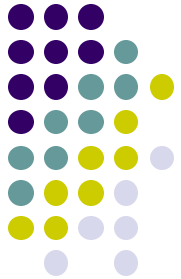


- describes the probability distribution of one random variable given the value of another.

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \qquad f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y) = f_{Y,X}(y|x) f_X(x)$$

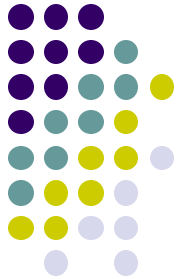
Bayesian Theorem and inference



- is a foundational concept in probability that allows us to update the probability of a hypothesis based on new evidence (like data).

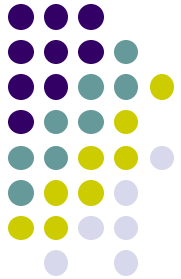
$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Sampling from Posterior



- Methods include MCMC, variational inference, Rejection Sampling.

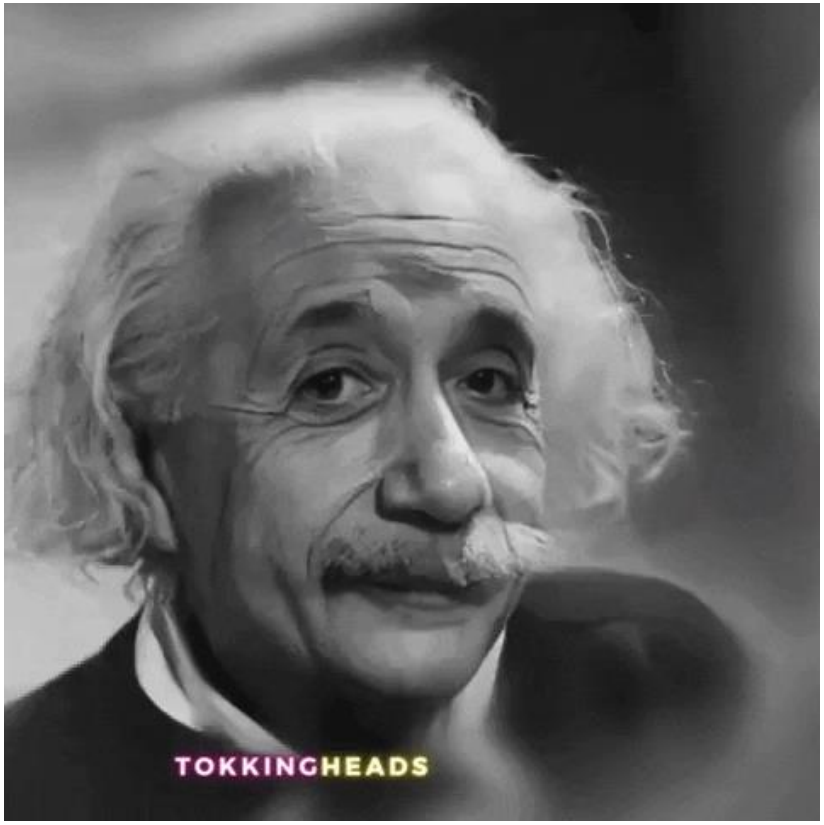
Summary



Key takeaways:

- Probability distributions are essential.
- Bayesian inference updates beliefs.
- Sampling techniques help in inference.

Elbert Einstien




Thank You
30 17:50