CS 540 Introduction to Artificial Intelligence

Statistics and Linear Algebra

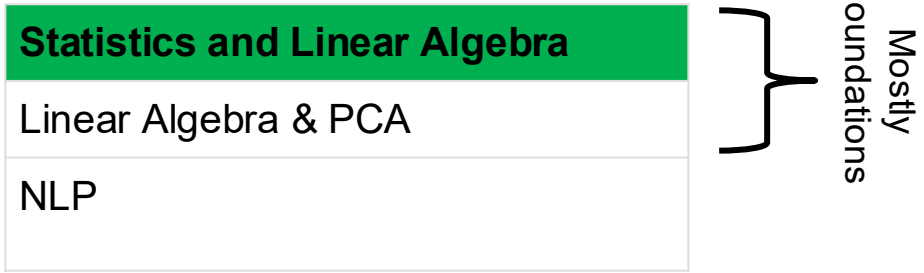
University of Wisconsin-Madison

Fall 2026 Sections 1 & 2

Announcements

- **HW 1 will be released tomorrow**:
  - Due **Friday Sep 18 at 11:59PM**
- TA discussion (optional) – review session every Tuesday at 4:00 PM in Morgridge Hall 3610

- Class roadmap:



Outline

- Probability Review
- Statistics
- Linear Algebra

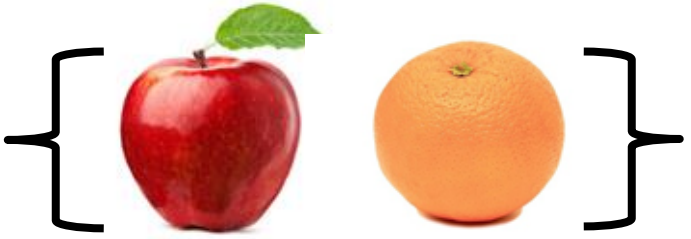
Review: **Random Variables**

- Intuitively: a number X that's random
- Mathematically:

  function that maps random outcomes to real values

$$X : \Omega \to \mathbb{R}$$

- Why?
  - Previously, everything is a set.
  - Real values are easier to work with



Review: **Joint Distributions**

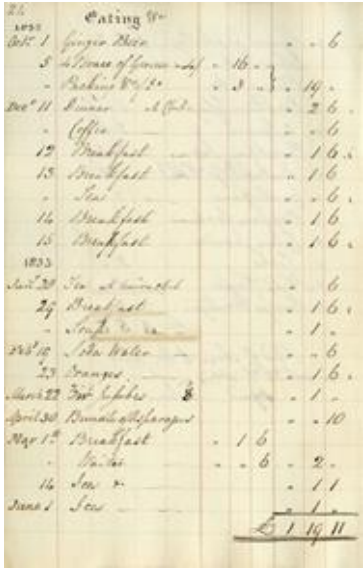
- Move from one variable to several
- Joint distribution: $P(X = a, Y = b)$
  - Why? Work with **multiple** types of uncertainty that correlate with each other

Review: **Marginal** Probability

- Given a joint distribution $P(X = a, Y = b)$
  - Get the distribution in just one variable:

$$P(X = a) = \sum_b P(X = a, Y = b)$$

  - This is the “marginal” distribution.



Review: Conditional Probability

For when **we know something** (i.e. Y=b)

$$P(X = a|Y = b) = \frac{P(X = a, Y = b)}{P(Y = b)}$$

	green	white
hot	150/365	45/365
cold	50/365	120/365

$$P(cold|white) = \frac{P(cold,white)}{P(white)} = \frac{120}{45+120} = 0.73$$

Review:Bayes' Rule

Theorem: For any events A and B we have

$$P(A|B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Proof: Apply the chain rule two different ways:

$$\left.\begin{aligned} P(A,B) &= P(A \mid B) \cdot P(B) \\ &= P(B \mid A) \cdot P(A) \end{aligned}\right\} \quad P(A|B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Review: Bayesian Inference

- Conditional Probability & Bayes Rule:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- Evidence *E*: what we can observe
- Hypothesis *H*: what we'd like to infer from evidence
  - Need to plug in prior, likelihood, etc.
- Usually do not know these probabilities, but can estimate them from data.

Correlation vs. Causation

- Conditional probabilities only define correlation (aka association)
- P(Y|X) “large” does not mean X causes Y
- Example: X=yellow finger, Y=lung cancer
- Common cause: smoking

Statistics

Samples and Estimation

- Usually, we don't know the distribution P
  - Instead, we see a bunch of samples

- Typical statistics problem: **estimate distribution** from samples
  - Estimate probabilities *P*(*H*), *P*(*E*), *P*(*E*|*H*)
  - Estimate the mean $E[X]$
  - Estimate parameters of a model $P_\theta(X)$

Samples and Estimation

- Example: Bernoulli with parameter p

  (i.e. a weighted coin flip)

  – $P(X = 1) = p$

  – Mean $E[X]$ is p

Examples: Sample Mean

- Bernoulli with parameter *p*
- See samples $x_1, x_2, \ldots, x_n$
  - Estimate mean with **sample mean**

$$\hat{\mathbb{E}}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - That is, counting heads

Estimating Multinomial Parameters

- *k*-sized die (special case: *k*=2 coin)
- Face *i* has probability p_i, for *i=1…k*
- In *n* rolls, we observe face *i* showing up n_i times

$$\sum_{i=1}^{k} n_i = n$$

- Estimate ($p_{1,\ldots,}\ p_k$) from this data ($n_{1,\ldots,}\ n_k$)

Maximum Likelihood Estimate (MLE)

- The MLE of multinomial parameters $(\widehat{p_1}, \dots, \widehat{p_k})$

$$\widehat{p_i} = \frac{n_i}{n}$$

- Estimate using frequencies

Regularized Estimate

- Hyperparameter $\epsilon > 0$

$$\widehat{p}_i = \frac{n_i + \epsilon}{n + k\epsilon}$$

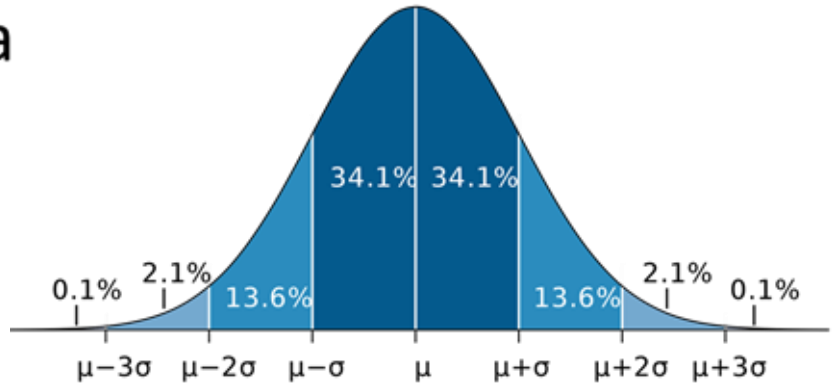
- Avoids zero when *n* is small
- Biased, but has smaller variance
- Also called smoothing, or the Maximum A Posteriori (MAP) estimate.

Estimating 1D Gaussian Parameters

- Gaussian (aka Normal) distribution $N(\mu, \sigma^2)$
  - True mean μ, true variance σ^2
- Observe n data points from this distribution

$$x_1, \ldots, x_n$$

- Estimate μ, σ^2 from this data



Wikipedia: Normal distribution

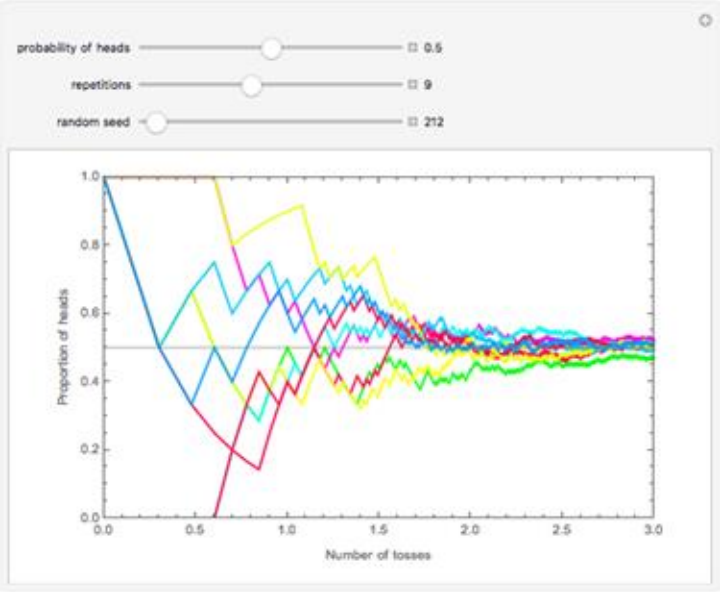
Estimating 1D Gaussian Parameters

- Mean estimate via sample mean $\hat{\mu} = \frac{x_1 + \cdots + x_n}{n}$
  This is an unbiased estimate of the mean

- Variance estimates
  - Unbiased $s^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n-1}$
  - MLE $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}$

Estimation Theory

- Is the sample mean a good estimate of the true mean?
  - Law of large numbers
  - Central limit theorems



Wolfram Demo

Estimation Errors

- With finite samples, likely error in the estimate.
- Mean squared error
  - $\mathrm{MSE}[\hat{\theta}] = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$
- Bias / Variance Decomposition
  - $\mathrm{MSE}[\hat{\theta}] = \mathbb{E}\left[(\hat{\theta} - E[\hat{\theta}])^2\right] + (\mathbb{E}[\hat{\theta}] - \theta)^2$

Variance Bias

Probability vs Statistics

- Probability reasons forward: given a model, work out what the data will look like.
  e.g: the coin is fair, what's the chance of 8 heads in 10 flips.

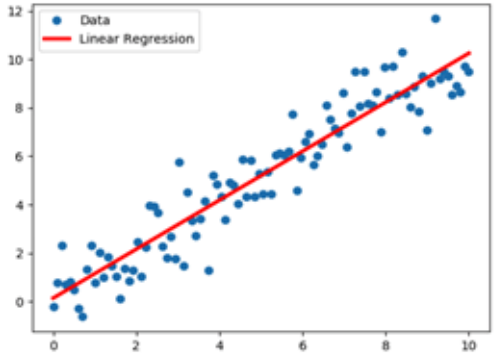
- Statistics reasons backward: given data, work out what the model was.
  e.g: I saw 8 heads in 10 flips — is this coin fair?.

- In machine learning:
  - Training a model is a statistics problem (data in, model out).
  - Using a trained model is a probability problem (model in, predictions out)

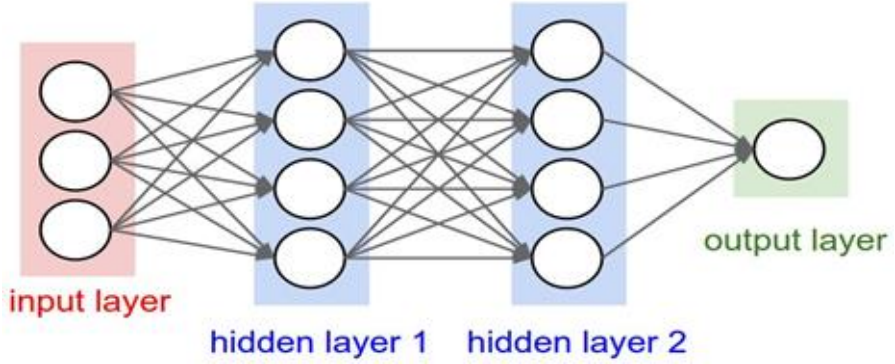
Linear Algebra

Linear Algebra

- Study of Linear functions: simple, tractable
- In AI/ML: building blocks for **all models**
  - e.g., linear regression; part of neural networks



Hieu Tran

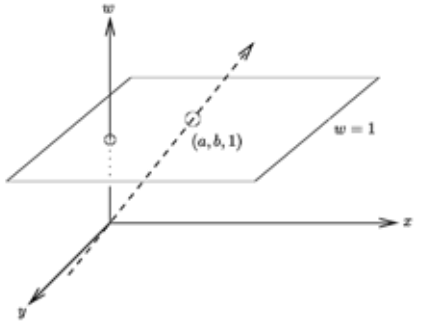


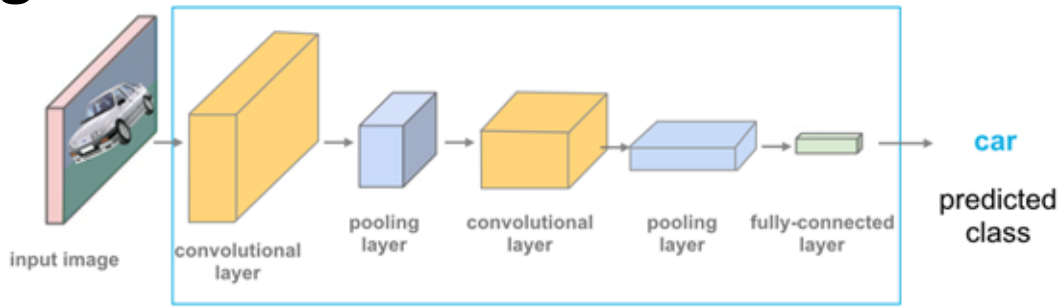
Stanford CS231n

Basics: **Vectors**

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \in \mathbb{R}^5$$

- Many interpretations
  - List of values (represents information)
  - **Point in space**
- Dimension: number of values: $x \in \mathbb{R}^d$
- AI/ML: often use very high dimensions:
  - Ex: images!





Basics: **Matrices**

- Many interpretations
  - Table of values; list of vectors
  - Represent linear transformations
  - Apply to a vector, get another vector

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{33} & A_{33} \\ A_{41} & A_{43} & A_{43} \end{bmatrix}$$

- Dimensions: # rows × # columns, $A \in \mathbb{R}^{m \times n}$
  - indexing

Basics: **Transposition**

- Transposes: flip rows and columns
  - Vector: standard is a column. Transpose: row vector
  - Matrix: go from $m \times n$ to $n \times m$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad x^T = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$$

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{bmatrix} \quad A^T = \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \\ A_{13} & A_{23} \end{bmatrix}$$

Matrix & Vector **Operations**

- **Vectors**
  - **Addition:** component-wise
    - Commutative: $x + y = y + x$
    - Associative: $(x + y) + z = x + (y + z)$

$$x + y = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \end{bmatrix}$$

  - **Scalar Multiplication**
    - Uniform stretch / scaling

$$cx = \begin{bmatrix} cx_1 \\ cx_2 \\ cx_3 \end{bmatrix}$$

Matrix & Vector **Operations**

- **Vector products**
  - **Inner product** (e.g., dot product)

$$< x, y >:= x^T y = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1 y_1 + x_2 y_2 + x_3 y_3$$

  - **Outer product**

$$xy^T = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & x_1 y_3 \\ x_2 y_1 & x_2 y_2 & x_2 y_3 \\ x_3 y_1 & x_3 y_2 & x_3 y_3 \end{bmatrix}$$

Matrix & Vector **Operations**

- x and y are **orthogonal** if $\langle x, y \rangle = 0$.
- Vector **norms**: "length"

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

- A set of vectors $\{x_1, x_2, \ldots x_n\}$ is **orthonormal** if:
  - For all pairs x_i, x_j we have $\langle x_i, x_j \rangle = 0$
  - For all x_i, we have $\|x\|_2 = 1$

Matrix & Vector **Operations**

- **Matrices:**
  - **Addition:** Component-wise
  - Commutative, Associative

$$A + B = \begin{bmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} + B_{21} & A_{22} + B_{22} \\ A_{31} + B_{31} & A_{32} + B_{32} \end{bmatrix}$$

  - **Scalar Multiplication**
  - "Stretching" the linear transformation

$$cA = \begin{bmatrix} cA_{11} & cA_{12} \\ cA_{21} & cA_{22} \\ cA_{31} & cA_{32} \end{bmatrix}$$

Matrix & Vector **Operations**

- **Matrix-Vector multiplication:**
  - **Linear transformation; plug in vector, get another vector**
  - Each entry in Ax is the inner product of a row of A with x

$$x \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}$$

$$Ax = \begin{bmatrix} \langle A_{1:}, x \rangle \\ \langle A_{2:}, x \rangle \\ \vdots \\ \langle A_{m:}, x \rangle \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n \\ \vdots \\ A_{m1}x_1 + A_{m2}x_2 + \cdots + A_{mn}x_n \end{bmatrix}$$

Matrix & Vector **Operations**

Ex: feedforward neural networks. Input *x*.

- Output of layer *k* is

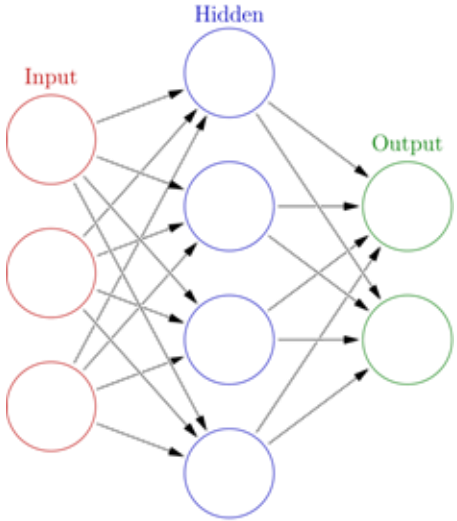
nonlinearity

$$f^{(k)}(x) = \sigma(W_k^T f^{(k-1)}(x)))$$

Output of layer k: vector

Weight **matrix** for layer k:
Note: linear transformation!

Output of layer k-1: **vector**

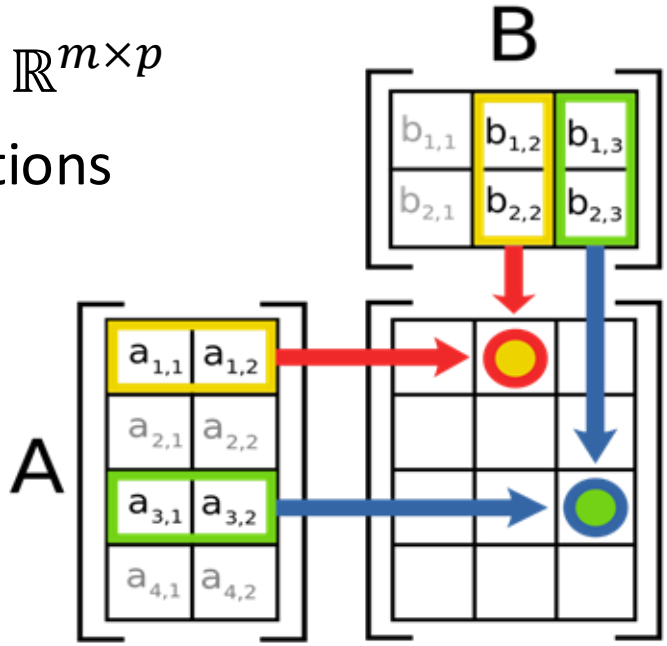


Wikipedia

Matrix & Vector **Operations**

- Matrix multiplication
  - $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, then $AB \in \mathbb{R}^{m \times p}$
  - "Composition" of linear transformations
  - Not commutative in general!

$$AB \neq BA$$



Wikipedia

Identity Matrix

- Like "1"
- Multiplying by it gets back the same matrix or vector

- Rows & columns are the "**standard basis vectors**" e_i

$$I = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

e_1 e_2 e_n

Readings

- Local classes: Math/Stat 431
- **Suggested reading:**
  - Probability and Statistics: The Science of Uncertainty, Michael J. Evans and Jeff S. Rosenthal
    http://www.utstat.toronto.edu/mikevans/jeffrosenthal/book.pdf
    (Chapters 1-3, excluding "advanced" sections)
  - Textbook: Artificial Intelligence: A Modern Approach (4th edition). Stuart Russell and Peter Norvig. Pearson, 2020. Appendix A