Short review of linear algebra, statistics, and probabilities

- **Based on chapters 2 and 3 of “Deep Learning”**

  http://www.deeplearningbook.org/

Linear algebra

- **Scalar: a single value.**

$$a \in \mathbb{R}, a \in \mathbb{N} \qquad a = 3$$

- **Vector: an array of values.**

$$a \in \mathbb{R}^D, a \in \mathbb{N}^D \qquad a = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix}$$

- **Matrix: a table of values.**

$$A \in \mathbb{R}^{D_1 \times D_2}, A \in \mathbb{N}^{D_1 \times D_2} \qquad A = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 2 & 9 \end{bmatrix}$$

Indexing notation

- **Indexing elements of a vector:** a_i

$$a = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix} \leftarrow a_1$$

Convention:
The first element
is the zero'th.

- **Indexing elements of a matrix:** a_{ij}

$$A = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 2 & 9 \end{bmatrix}$$

a_{12}

Simple operations

- **Transpose**

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$$

$$\mathbf{a}^\top = \begin{bmatrix} a_0 & a_1 & a_2 \end{bmatrix}$$

$$(\mathbf{A}_{ij})^\top = \mathbf{A}_{ji}$$

- **Addition**
  - **Vectors and matrices w. the same shape**

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \qquad \mathbf{a} + \mathbf{b} = \begin{bmatrix} a_0 + b_0 \\ a_1 + b_1 \\ a_2 + b_2 \end{bmatrix}$$

$$(\mathbf{A} + \mathbf{B})_{ij} = \mathbf{A}_{ij} + \mathbf{B}_{ij}$$

Simple operations

- Multiply by a scalar $\alpha \mathbf{a} = \begin{bmatrix} \alpha\ a_0 \\ \alpha\ a_1 \\ \alpha\ a_2 \end{bmatrix}$
- Vector product.
  - The dot product $\mathbf{a}^\top \mathbf{a} = \sum_i a_i a_i$
    - Note: it yields a scalar.
  - Element-wise product: $\mathbf{a} \odot \mathbf{a} = \begin{bmatrix} a_0 a_0 \\ a_1 a_1 \\ a_2 a_2 \end{bmatrix}$
    - Also known as Hadamard product

Operations

- **Matrix product (dot product):**

$$C_{ij} = \sum_k A_{ik} B_{kj}$$

  - **A's columns must equal B's rows (order is important)**

$$A \in \mathbb{R}^{D_1 \times D_2}, B \in \mathbb{R}^{D_2 \times D_3}$$

  - **Distributive:** $A(B + C) = AB + AC$
  - **Associative:** $A(BC) = (AB)C$
  - **Product of transpose:** $(AB)^\top = B^\top A^\top$

Inverse

- We denote a matrix's inverse as A^{-1}
- A matrix has an inverse iff:
  - it's square. $D_1 = D_2$
  - its columns are linearly independent.
    - No column can be recovered using a combination of other columns

A square matrix not invertible is *singular*

- Inverses are useful to solve systems of equations:

$$Ax = b \qquad x = A^{-1}b$$

Norms

- L^p norm. Size of a vector (or matrix)

$$\| a \|_p = \left(\sum_i |a_i|^p \right)^{1/p}$$

- Standard norms in ML:
  - Euclidean norm (p=2) $\| a \|_2 = \sqrt{\left(\sum_i |a_i|^2 \right)}$
  - Dot product w. 2-norm: $a^\top b = ||a||_2 ||b||_2 \cos \theta_{ab}$
  - Frobenius norm (matrix): $\| A \|_2 = \sqrt{\left(\sum_i \sum_j |a_{ij}|^2 \right)}$

Special matrices & vectors

- Identity. Denoted I_n.

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

  - All zeros except for ones on the main diagonal.
- Symmetric: $A = A^\top$
- Unit vector: $\|a\|_2 = 1$
- Orthogonal vectors: $a^\top b = 0$
- Orthonormal vectors: unit and orthogonal $A^\top A = AA^\top = I$
- Orthogonal matrix: Orthonormal rows & columns

- Skip eigendecomposition, SVD, pseudo-Inverse, determinants (Sections 2.7–2.11).
  - We will get back to them if/when needed in the course.

- On to probabilities
- Chapter 3 of “Deep Learning”
  - I’ve adapted some of the lecture slides from the book.
  - Thanks to Ian Goodfellow for providing slides.

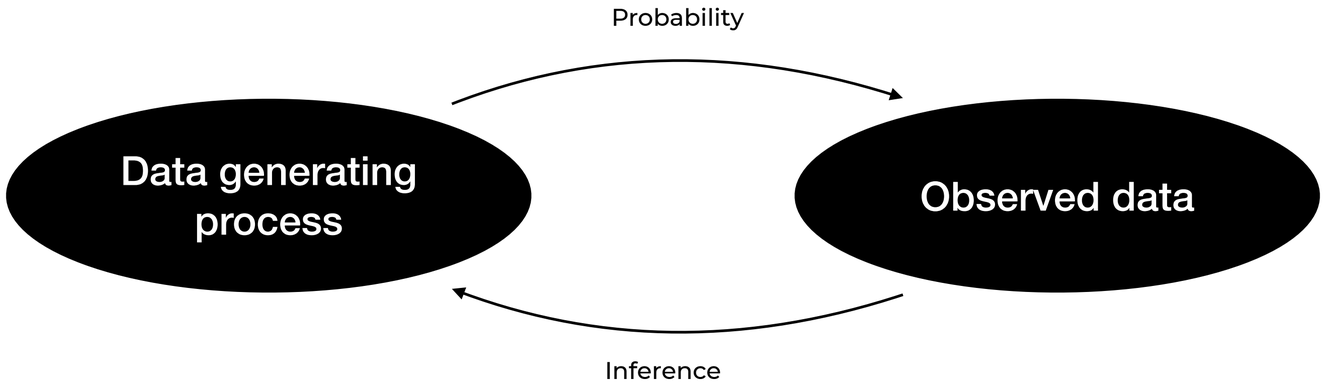
Why probabilities?

- To capture uncertainty

  E.g., What time will I get home tonight?

- Probabilities provide a formalism for making statements about “data generating processes” (L. Wasserman)

  E.g., what happens when I flip a fair coin?



The 🎲 example

- Generate data by throwing a fair die.
- What do we know about a single throw?
  - 6 possible outcomes. (sample space)
  - Each outcome (e.g., 1). (element, state)
    - A subset of outcomes (e.g., <3). (event)
  - Outcomes are equiprobable. (uniform distribution)

Random variables and probabilities

- A random variable (r.v.) is a probabilistic outcome.
  - For example,
    - Die throw (X)
    - The actual outcome is $\in \{1, 2, 3, 4, 5, 6\}$. (x)
- A probability function (P) assigns a real number to each possible event: $P(x) \geq 0, \forall x \in X$

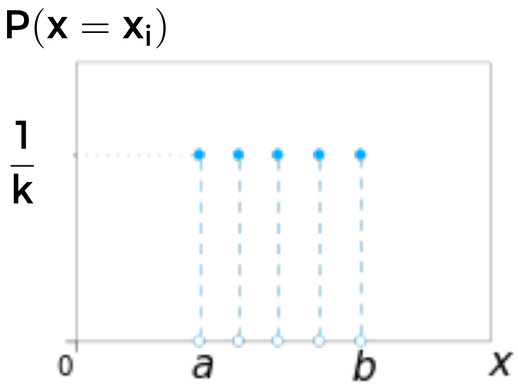
$$P(\bigcup x) = 1$$

Discrete RVs

- An RV is discrete if it takes a finite number of values[1]

$$P(x = x_i) \geq 0, \forall i$$

$$\sum_i P(x = x_i) = 1$$

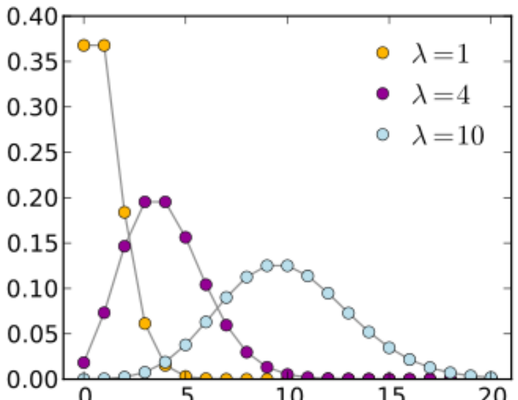


- E.g., uniform distribution:

$$P(x = x_i) = \frac{1}{k}, \forall i$$

- E.g., Poisson distribution:

$$P(x = x_i; \lambda) = \frac{\lambda^{x_i} \exp^{-\lambda}}{x_i!}$$



both images are from: wikipedia.org

1. technically: it must be countable

Continuous RVs

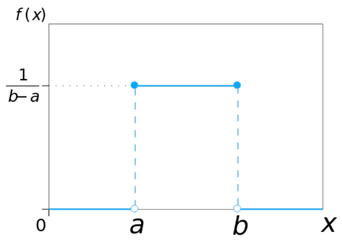
- An RV is continuous if $f(x) \geq 0, \forall x \in X$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(a < x < b) = \int_{a}^{b} f(x)dx$$

- f(x) is a probability density function (PDF)
  - E.g., (continuous) uniform distribution:

$$u(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$



from: wikipedia.org

  - E.g., Gaussian distribution

A few useful properties
(shown for discrete variables for simplicity)

- **Sum rule:** $P(X) = \sum_Y P(X, Y)$
- **Product rule:** $P(X, Y) = P(X \mid Y)P(Y)$
- **Chain rule:** $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})P(X_1)$
- **If x and y are independent:** $P(X, Y) = P(X)P(Y)$
- **Bayes' Rule:** $P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$

Moments

- **Expectation:** $\mathbb{E}[X] = \sum_i P(x = x_i) x_i \qquad \mathbb{E}[aX] = a\mathbb{E}[X]$
- **Variance:** $\sigma^2 = E[(X - E[X])^2]$
- **Covariance:** $Cov(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$
  - **correlation:** $\rho(x, y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$

Further Reading

- Prologue to “The Master Algorithm”
  http://homes.cs.washington.edu/~pedrod/Prologue.pdf
- Ch. 1 of Hastie et al.
- Math Preparation
  - Ch.2 of Pattern Recognition and Machine Learning [PRML]
  - Ch.2-3 of Deep Learning [DL]
  - Slightly more advanced:

    http://www.cs.mcgill.ca/~dprecup/courses/ML/Materials/prob-review.pdf

    http://www.cs.mcgill.ca/~dprecup/courses/ML/Materials/linalg-review.pdf