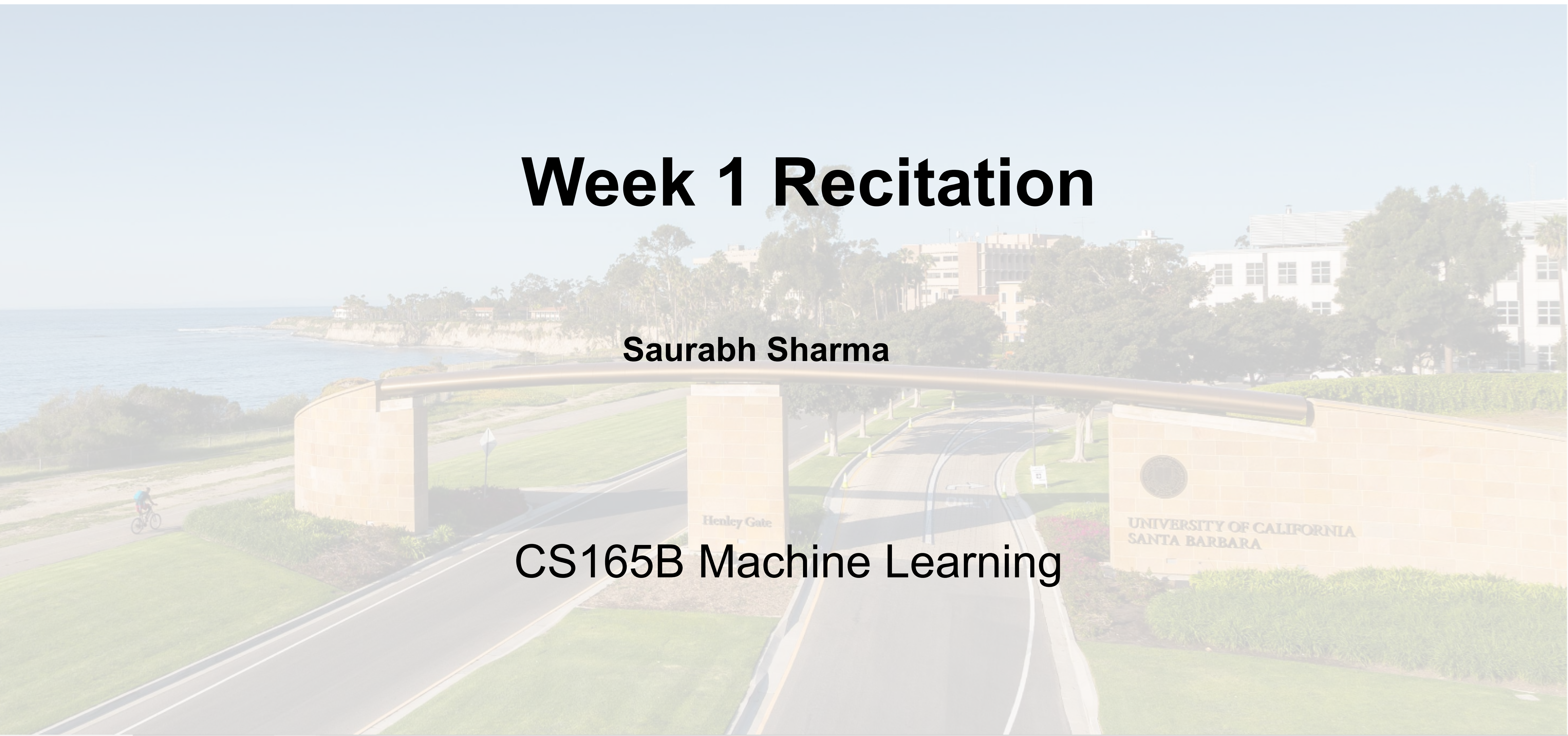
Week 1 Recitation

Saurabh Sharma

CS165B Machine Learning



Outline for today

- Recap of probability concepts

- Maximum likelihood estimation

- Recap of linear algebra concepts

Random Variable (RV)

- A random variable is a variable that can take on different values randomly.

- A probability distribution is a description of how likely a random variable or set of random variables is to take on each of its possible states.

Random Variable (RV)

- Example

Sample space (outcomes)	Number of Heads (Random Variable X)	How many possible ways can this happen, if you flip a coin twice? (Probability of X)
HH	2	1 out of 4 outcomes
HT	1	2 out of 4 outcomes
TH		
TT	0	1 out of 4 outcomes

Value of X	2	1	0
Probability of X: $p(x)$ or $f(x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

Discrete random variable

- The probability distribution of a discrete RV is given by its probability mass function, or PMF.

- To be a probability mass function on a random variable $\mathbf{x}$, a function P must satisfy the following properties:

    - The domain of P must be the set of all possible states of $\mathbf{x}$.

    - $\forall x \in X, 0 \leq P(x) \leq 1.$

    - $\Sigma_{x \in X} P(x) = 1$

Continuous random variable

- To be a probability density function, a function p must satisfy the following properties:
  - The domain of p must be the set of all possible states of x.
  - $\forall x \in X, p(x) \geq 0$. Note that we don't require $p(x) \leq 1$.
  - $\int p(x)dx = 1$
- We can integrate the density function to find the actual probability mass of a set of points.
- Probability that x lies in the interval $[a, b]$ is given by $\int_{[a,b]} p(x)dx$.

Joint and marginal probability

- A probability distribution over multiple RVs is known as a *joint probability distribution*.
- $P(\mathbf{x} = x, \mathbf{y} = y)$ denotes the probability that $\mathbf{x} = x$ and $\mathbf{y} = y$ simultaneously.
- For discrete RVs, given the joint distribution $P(x, y)$, we can get the *marginal distribution* $P(x)$ by the sum rule:

$$P(\mathbf{x} = x) = \Sigma_y P(\mathbf{x} = x, \mathbf{y} = y)$$

- For continuous RVs,

$$p(x) = \int p(x, y)$$

Conditional probability

- Probability of an event given that another event has happened:

$$P(\mathbf{y} = y \mid \mathbf{x} = x) = \frac{P(\mathbf{y} = y, \mathbf{x} = x)}{P(\mathbf{x} = x)}$$

- *Chain rule* of probability:

$$P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(\mathbf{n})}) = P(\mathbf{x}^{(1)}) \Pi_{i=2}^{n} P(x^{(i)} \mid x^{(1)}, \ldots, x^{(i-1)})$$

- The rule follows directly from the definition of conditional probability:

$$\begin{aligned} P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c}) \\ P(\mathrm{b}, \mathrm{c}) &= P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \\ P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}). \end{aligned}$$

Independence

- Two random variables $\mathbf{x}$ and $\mathbf{y}$ are *independent* if their probability distribution can be expressed as,

$$P(\mathbf{x} = x, \mathbf{y} = y) = P(\mathbf{x} = x)P(\mathbf{y} = y)$$

Expectation and Variance

- The *expectation* or *expected value* of some function $f(x)$ with respect to a probability distribution $P(x)$ is the average or mean value that f takes on when x is drawn from P.

- For discrete RVs,

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_x P(x) f(x)$$

- For continuous RVs,

$$\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x) f(x) dx$$

Expectation and Variance

- Expectations are linear,

$$\mathbb{E}_{\mathbf{x}}[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_{\mathbf{x}}[f(x)] + \beta \mathbb{E}_{\mathbf{x}}[g(x)]$$

- The *variance* gives a measure of how much the values of a function of a random variable **x** vary as we sample different values of **x** from its probability distribution:

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$$

  - The square root of the variance is called the standard deviation.

Useful probability distributions

- The *Bernoulli* distribution is a distribution over a single binary random variable. It is controlled by a single parameter $\phi \in$ [0,1], which gives the probability of this variable being equal to 1:

$$P(\mathrm{x} = 1) = \phi$$

$$P(\mathrm{x} = 0) = 1 - \phi$$

- The *Multinoulli* distribution extends the above to the case when x can take k states. It is controlled by $k - 1$ parameters that specify the probabilities at the k states.

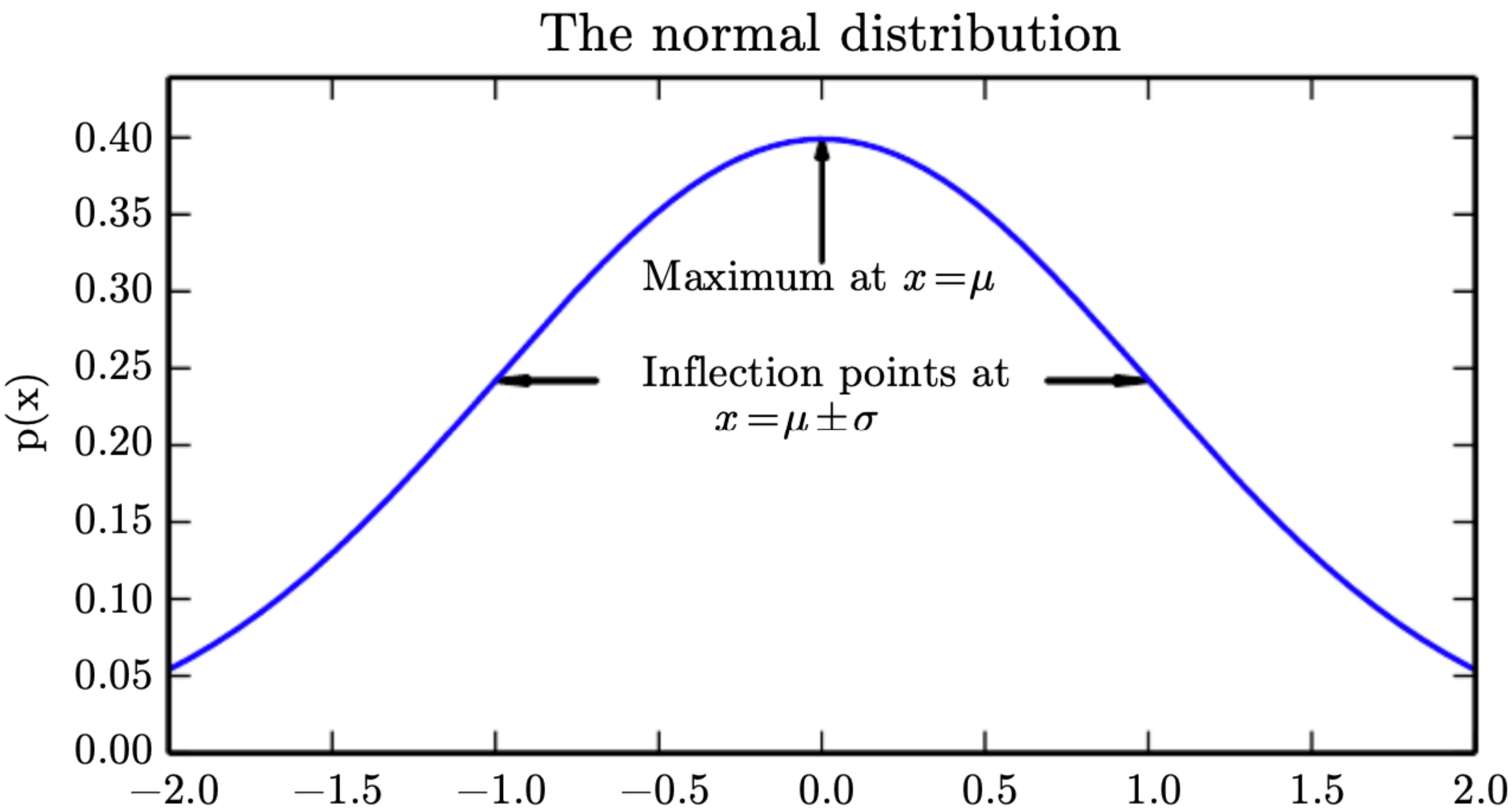
Useful probability distributions

- The most commonly used distribution over real numbers is the normal distribution, also known as the *Gaussian* distribution:

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- The expectation of this distribution is μ and it's variance is σ^2.

Useful probability distributions



- μ gives the location of the central peak and σ controls the width of the peak. In the above plot, $\mu = 0, \sigma = 1$.

Bayes Rule

- Sometimes we know $P(y \mid x)$ and need to compute $P(x \mid y)$. In this case, if we also know $P(x)$, we can use Bayes' rule,

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

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

$$P(x \mid y) = \frac{P(x)P(y \mid x)}{\Sigma_{x' \in X} P(x')P(y \mid x')}$$

Maximum likelihood estimation

Example: $X_1,\dots,X_n$ - i.i.d. random variables with probability $p_X(x|\theta) = P(X=x)$ where θ is a parameter

- likelihood function $L(\theta|x)$ where $x=(x_1,\dots,x_n)$ is set of observations

$$L(\theta \mid x) = \prod_{i=1}^{n} p_X(x_i \mid \theta)$$

- maximum likelihood estimate $\hat{\theta}(x)$

maximizer of $L(\theta|x)$

- typically easier to work with log-likelihood function, $\ell(\theta|x) = \log L(\theta|x)$

MLE example

- Suppose we have three data points they have been generated from a process that is adequately described by a Gaussian distribution. These points are 9, 9.5 and 11.

- *How do we calculate the maximum likelihood estimates of the parameter values of the Gaussian distribution μ and σ?*

- The Gaussian density function is given by,

$$P(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

MLE example

- The *likelihood* or the joint density of the data is given by,

$$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

- The *log-likelihood* is given by,

$$\ln(P(x; \mu, \sigma)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9.5-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(11-\mu)^2}{2\sigma^2}$$

MLE example

- This can be further simplified to,

$$\ln(P(x;\mu,\sigma)) = -3\ln(\sigma) - \frac{3}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\left[(9-\mu)^2 + (9.5-\mu)^2 + (11-\mu)^2\right]$$

- This expression attains it's maximum for μ when the partial derivative with respect to μ is 0,

$$\frac{\partial \ln(P(x;\mu,\sigma))}{\partial \mu} = \frac{1}{\sigma^2}[9 + 9.5 + 11 - 3\mu] = 0$$

- $$\mu = \frac{9 + 9.5 + 11}{3} = 9.833$$

Recap of linear algebra concepts

Scalars and Vectors

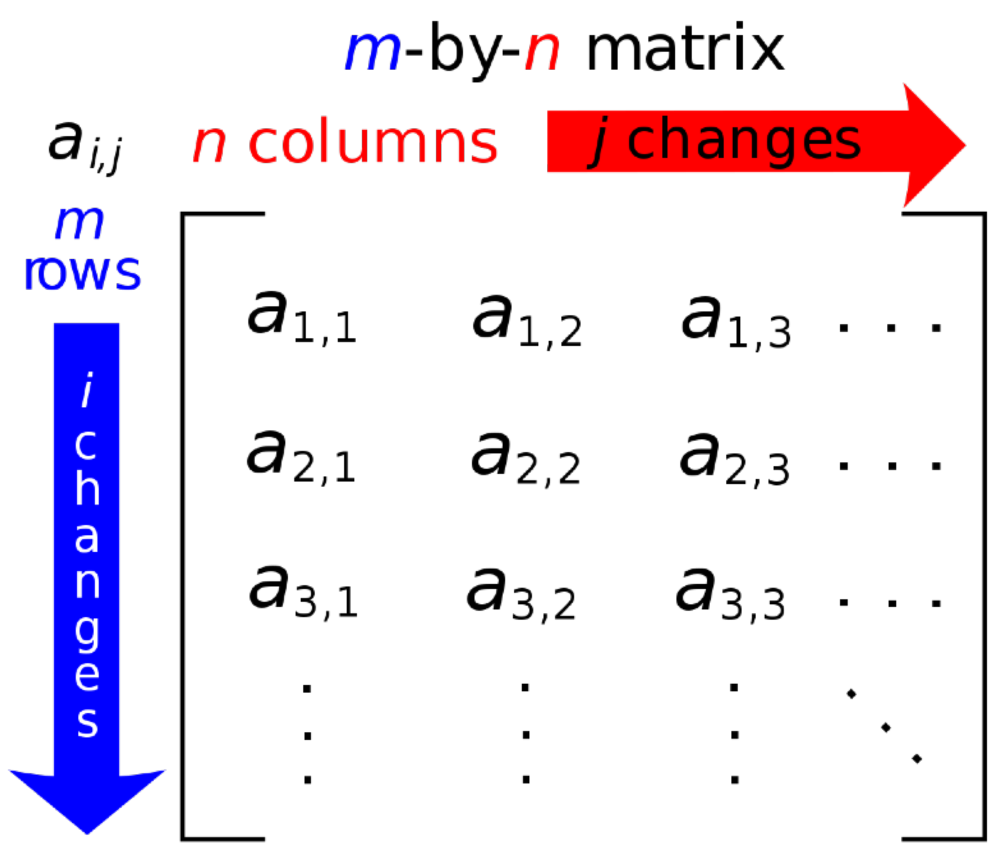
- *Scalars*: A scalar is just a single number. Example: 5, 10, 15

- *Vectors*: A vector is an ordered array of numbers. We can identify each individual number by its index in that ordering. Example:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Matrices and tensors

- *Matrices*: A matrix is a rectangular array of numbers, and we can identify each number using its *row and column indices*.



- *Tensors*: A tensor is like a high-dimensional matrix that can be indexed similarly. For example, the element at (i, j, k) coordinate of a 3D tensor $\mathbf{A}$ is denoted by $\mathbf{A}_{i,j,k}$.

Matrix and vector operations

- *Matrix addition:* Matrices can be added as long as their shapes match

$$C = A + B, \text{ where } C_{i,j} = A_{i,j} + B_{i,j}$$

- *Scalar multiplication and addition:* Scalars can be multiplied and added to each element of a matrix

$$D = a \cdot B + c, \text{ where } D_{i,j} = a \cdot B_{i,j} + c$$

- *Broadcasting:* Vectors can be added to matrices (shapes must match)

$$C = A + B, \text{ where } C_{i,j} = A_{i,j} + B_j$$

The vector B gets added to every row in A.

Matrix and vector operations

- *Dot product* of two vectors $x^T y = y^T x = \Sigma_k x_k y_k$
- *Product of two matrices* $C = AB$ is defined as

$$C_{i,j} = \Sigma_k A_{i,k} B_{k,j}$$

$C_{i,j}$ is the *dot product* of the ith row of A and jth column of B.
Number of columns in A must match number of rows in B.

- *Distributive law:* $$A(B + C) = AB + AC$$
- *Associativity:* $$A(BC) = (AB)C$$
- *Commutativity* does not always hold: $$AB \neq BA$$

Matrix and vector operations

- *Transpose* of a matrix is obtained by “flipping” along the diagonal.

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \Rightarrow \boldsymbol{A}^\top = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

- *Transpose of a product:*

$$(AB)^T = B^T A^T$$

Some special matrices

- *A square matrix* has the same number of rows and columns. *The identity matrix* is a square matrix with 1s along the diagonal:

$$I_1 = [1],\ I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\ \ldots,\ I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

- *The inverse of a square matrix* A is a matrix A^{-1} that satisfies:

$$AA^{-1} = A^{-1}A = I$$

Norms

- *Norm* is a function that intuitively measures the size of a vector.
- L^1 norm : $\|x\|_1 = \Sigma_i |x_i|$
- L^2 norm : $\|x\|_2 = \sqrt{\Sigma_i |x_i|^2}$
- L^∞ norm : $\|x\|_\infty = \max_i |x_i|$. Also known as the max norm.
- We can also assign a norm to a matrix. The most commonly used is the *Frobenius norm*: $\|A\|_F = \sqrt{\Sigma_{i,j} |A_{i,j}|^2}$.
  - This is analogous to the L_2 norm of a vector.

References

- **Probability and linear algebra recap:** Chapters 2 and 3 of the Deep learning book (GBC).

- **Maximum likelihood estimation:** https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1