Math Review

David Robinson

Vector

A **vector** is a one-dimensional array of ordered real-valued scalars.

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

Norm of a Vector

A vector **norm** is a function that maps a vector to a scalar value, measuring the size of the vector. The norm, f, should satisfy the following properties.

1. Scaling: $f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$
2. Triangle inequality: $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$
3. Must be non-negative: $f(\mathbf{x}) \geq 0$

General equation for the norm of a vector

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} \|x_i\|^p \right)^{\frac{1}{p}}$$

Most Common Norms

1. L1 Norm
$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$
2. L2 Norm
$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$$
3. Max Norm
$$\|\mathbf{x}\|_\infty = \max |x_i|$$

Vector Projection

Orthogonal projection of a vector $\mathbf{y}$ onto vector $\mathbf{x}$ is represented by

$$\mathbf{proj}_{\mathbf{x}}(\mathbf{y}) = \frac{\mathbf{x} \cdot \|\mathbf{y}\| \cdot \cos(\theta)}{\|\mathbf{x}\|}$$

Hyperplanes

A **hyperplane** is a subspace whose dimension is one less than that of its ambient space. For example, a hyperplane in a 2D space is one-dimensional and a hyperplane in a 3D space is two-dimensional. Hyperplanes are decision boundaries used for linear classification.

Matrices

A **matrix** is a rectangular array of real-valued scalars arranged in m horizontal rows and n vertical columns.

Gradient

The **gradient** of the multivariate function $f(\mathbf{x})$ with respect to the n-dimensional input vector $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$, is a vector of n partial derivatives.

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

The gradient descent algorithm relies on the opposite direction of the gradient of the loss function $\mathcal{L}$ with respect to the model parameters $\theta(\nabla_\theta \mathcal{L})$ for minimizing the loss.

Stationary Points

Stationary points of a differentiable function $f(x)$ of one variable are the points where the derivative of the function is zero, such as a minimum or maximum.

Random Variables

A **probability distribution** is a description of how likely a random variable is to take on each of its possible states.

1. **Joint probability distribution** acts on many variables at the same time.
2. **Marginal probability distribution** acts on a single variable.
3. **Conditional probability distribution** is on one variable when another variable has taken a certain value.

Bayes' Theorem

Bayes' Theorem can calculate conditional probabilities for one variable when conditional probabilities for another variable are known.

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Independence

Two random variables are **independent** if the occurrence of one of the variables does not affect the occurrence of the other variable.

Expected Value

The **expected value** of a function $f(X)$ with respect to a probability distribution $P(X)$ is the mean when X is drawn from $P(X)$.

For a discrete random variable,

$$\mathbb{E}_{X \sim P}[f(X)] = \sum_X P(X) f(X)$$

For a continuous random variable,

$$\mathbb{E}_{X \sim P}[f(X)] = \int P(X) f(X)\, dX$$

Variance

Variance measures how much the values of the function $f(X)$ deviate from the expected value.

$$\sigma^2 = \mathbb{E}\Big[(f(X) - \mathbb{E}[f(x)])^2\Big]$$

Covariance

Covariance measures how much two random variables are linearly related to each other.

$$\text{Cov}(f(x), g(Y)) = \mathbb{E}\Big[(f(X) - \mathbb{E}[f(x)])(g(Y) - \mathbb{E}[g(Y)])\Big]$$

Correlation

The **correlation coefficient** is the covariance normalized by the standard deviations of the two variables.

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y}$$

Probability Distributions

1. **Bernoulli Distribution**

   The probability to be 1 is p and the probability to be 0 is $1 - p$.

2. **Uniform Distribution**

   The probability of each value $i \in \{1, 2, \ldots, n\}$ is $p_i = \frac{1}{n}$.

3. **Binomial Distribution**

   The probability of getting k successes in n trials is $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

4. **Poisson Distribution**

   A discrete random variable X with states $k \in \{0, 1, 2, \ldots\}$ and a number of events occurring independently in a fixed interval of time with a known rate λ has probability

   $$P(X = k) = \frac{\lambda^X \cdot e^{-\lambda}}{X!}$$

5. **Gaussian Distribution**

   For a random variable X with n independent measurements, the density is

   $$P_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$