Math Essentials

https://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/Mathematics%20for%20machine%20learning.pdf
https://courses.washington.edu/css490/2012.Winter/lecture_slides/02_math_essentials.pdf
https://gwthomas.github.io/docs/math4ml.pdf

Linear Algebra

See linalg.pdf

Calculus

List of Derivative Rules

- **Constant Rule:** $f(x) = c$ then $f'(x) = 0$
- **Constant Multiple Rule:** $g(x) = c \cdot f(x)$ then $g'(x) = c \cdot f'(x)$
- **Power Rule:** $f(x) = x^n$ then $f'(x) = nx^{n-1}$
- **Sum and Difference Rule:** $h(x) = f(x) \pm g(x)$ then $h'(x) = f'(x) \pm g'(x)$
- **Product Rule:** $h(x) = f(x)g(x)$ then $h'(x) = f'(x)g(x) + f(x)g'(x)$
- **Quotient Rule:** $h(x) = \dfrac{f(x)}{g(x)}$ then $h'(x) = \dfrac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$
- **Chain Rule:** $h(x) = f(g(x))$ then $h'(x) = f'(g(x))g'(x)$

https://www.math.ucdavis.edu/~kouba/Math17BHWDIRECTORY/Derivatives.pdf

List of Derivative Rules

- **Exponential Derivatives**
  - $f(x) = a^x$ then $f'(x) = \ln(a)a^x$
  - $f(x) = e^x$ then $f'(x) = e^x$
  - $f(x) = a^{g(x)}$ then $f'(x) = \ln(a)a^{g(x)}g'(x)$
  - $f(x) = e^{g(x)}$ then $f'(x) = e^{g(x)}g'(x)$

- **Logarithm Derivatives**
  - $f(x) = \log_a(x)$ then $f'(x) = \dfrac{1}{\ln(a)x}$
  - $f(x) = \ln(x)$ then $f'(x) = \dfrac{1}{x}$
  - $f(x) = \log_a(g(x))$ then $f'(x) = \dfrac{g'(x)}{\ln(a)g(x)}$
  - $f(x) = \ln(g(x))$ then $f'(x) = \dfrac{g'(x)}{g(x)}$

https://www.math.ucdavis.edu/~kouba/Math17BHWDIRECTORY/Derivatives.pdf

Partial Derivative

A **partial derivative** of a function of several variables is its derivative with respect to one of those variables, with the others held constant

$$\frac{\partial y}{\partial x}$$

Examples

- Derivative of sigmoid function $s(x) = \frac{1}{1+e^{-x}}$ $\rightarrow$ $\frac{ds}{dx} =?$
- $f(x) = ln(3x - 4)\ \frac{df}{dx} =?$
- $f(x) = ln[(1+x)(1+x^2)^2(1+x^3)^3]\ \frac{df}{dx} =?$
- $\frac{\partial}{\partial x}\ln(x^2 + y^2) =?$
- $\frac{\partial}{\partial y}\ln(x^2 + y^2) =?$

Probability & Statistics

Probability

- Data comes from a process that is not completely known.
- This lack of knowledge is indicated by modeling the process as a random process.
- Maybe the process is actually deterministic, but because we do not have access to complete knowledge about it, we model it as random and use probability theory to analyze it.
- Tossing a coin is a random process because we cannot predict at any toss whether the outcome will be heads or tail.

Probability

- The extra pieces of knowledge that we do not have access to are named the ***unobservable variables***.
- In the coin tossing example, the only ***observable variable*** is the outcome of the toss.
- Denoting the unobservables by $\boldsymbol{z}$ and the observable as x, in reality we have

$$x = f(\boldsymbol{z})$$

where $f(\cdot)$ is the deterministic function that defines the outcome from the unobservable pieces of knowledge.

Probability

- A random experiment is one whose outcome is not predictable with certainty in advance.
- The set of all possible outcomes is known as the *sample space S*.
- A sample space is ***discrete*** if it consists of a finite (or countably infinite) set of outcomes; otherwise it is ***continuous***.
- Any subset *E* of *S* is an ***event***.
- Events are sets, and we can talk about their complement, intersection, union, and so forth.

Probability

- One interpretation of probability is as a ***frequency***.
- When an experiment is continually repeated under the exact same conditions, for any event E, the proportion of time that the outcome is in E approaches some constant value.
- This constant limiting frequency is the probability of the event, and we denote it as $P(E)$.

Probability

Probability sometimes is interpreted as a ***degree of belief***.

- For example, when we speak of Turkey's probability of winning the World Soccer Cup in 2014, we do not mean a frequency of occurrence, since the championship will happen only once and it has not yet occurred (at the time of the writing of this book).

Axioms of Probability

$0 \leq P(E) \leq 1$.

- If E_1 is an event that cannot possibly occur, then $P(E_1) = 0$.
- If E_2 is sure to occur, $P(E_2) = 1$.

S is the sample space containing all possible outcomes, $P(S) = 1$.

Axioms of Probability

If E_i, i = 1, . . . , n are mutually exclusive (i.e., if they cannot occur at the same time, as in $E_i \cap E_j = \emptyset$, j = i, where ∅ is the ***null event*** that does not contain any possible outcomes), we have

$$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$$

For example, letting E^c denote the *complement* of *E*, consisting of all possible outcomes in *S* that are not in *E*, we have $E \cap E^c = \emptyset$ and

$$P(E \cup E^c) = P(E) + P(E^c) = 1 \quad \Rightarrow P(E^c) = 1 - P(E)$$

Axioms of Probability

- If the intersection of E and F is not empty, we have

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Conditional Probability

- Probability *P(E|F)* is the probability of the occurrence of event *E* given that *F* occurred and is given as

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

- Knowing that *F* occurred reduces the sample space to *F*, and the part of it where *E* also occurred is $E \cap F$.

Conditional Probability

Because ∩ is commutative, we have

$$P(E \cap F) = P(E|F)P(F) = P(F|E)P(E)$$

which gives us *Bayes' formula*:

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

Conditional Probability

When F_i are mutually exclusive and exhaustive, namely, $\bigcup_{i=1}^{n} F_i = S$

$$E = \bigcup_{i=1}^{n} E \cap F_i$$

$$P(E) = \sum_{i=1}^{n} P(E \cap F_i) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

Bayes' formula allows us to write

$$P(F_i|E) = \frac{P(E \cap F_i)}{P(E)} = \frac{P(E|F_i)P(F_i)}{\sum_j P(E|F_j)P(F_j)}$$

If E and F are *independent*, we have $P(E|F) = P(E)$ and thus

$$P(E \cap F) = P(E)P(F)$$

That is, knowledge of whether F has occurred does not change the probability that E occurs.

Random Variables

A ***random variable*** is a function that assigns a number to each outcome in the sample space of a random experiment.

Probability Distribution and Density Functions

The *probability distribution function* $F(\cdot)$ of a random variable X for any real number a is

$$F(a) = P\{X \le a\}$$

and we have

$$P\{a < X \le b\} = F(b) - F(a)$$

Probability Distribution and Density Functions

If *X* is a discrete random variable

$$F(a) = \sum_{\forall x \leq a} P(x)$$

where *P(·)* is the *probability mass function* defined as *P(a)* = *P*{*X* = *a*}.

If *X* is a *continuous* random variable, *p(·)* is the *probability density function* such that

$$F(a) = \int_{-\infty}^{a} p(x)dx$$

Joint Distribution and Density Functions

- In certain experiments, we may be interested in the relationship between two or more random variables, and we use the *joint* probability distribution and density functions of *X* and *Y* satisfying
  - $F(x, y) = P\{X \leq x, Y \leq y\}$
- Individual *marginal* distributions and densities can be computed by marginalizing, namely, summing over the free variable:

$$F_X(x) = P\{X \leq x\} = P\{X \leq x, Y \leq \infty\} = F(x, \infty)$$

Joint Distribution and Density Functions

- In the discrete case $P(X = x) = \sum_j P(x, y_i)$

- In the continuous case $p_X(x) = \int_{-\infty}^{\infty} p(x, y)dy$

Expected value

- *Expectation*, *expected value*, or *mean* of a random variable *X*, denoted by *E[X]*, is the average value of *X* in a large number of experiments:
  - $E[X] = \begin{cases} \sum_i x_i P(x_i) \text{ if } X \text{ is discrete} \\ \int xp(x)dx \text{ if } X \text{ is continous} \end{cases}$
- It is a weighted average where each value is weighted by the probability that *X* takes that value. It has the following properties $(a, b \in \Re)$

$E[aX + b] = aE[X] + b$

$E[X + Y] = E[X] + E[Y]$

Expected value

For any real-valued function *g(·)*, the expected value is

$$E[g(X)] = \begin{cases} \sum_i g(x_i)P(x_i) \text{ if } X \text{ is discrete} \\ \int g(x)p(x)dx \text{ if } X \text{ is continous} \end{cases}$$

A special $g(x) = x^n$, called the *n*th moment of *X*, is defined as

$$E[X^n] = \begin{cases} \sum_i x_i^n P(x_i) \text{ if } X \text{ is discrete} \\ \int x^n p(x)dx \text{ if } X \text{ is continous} \end{cases}$$

Mean is the first moment and is denoted by μ.

Variance

- *Variance* measures how much *X* varies around the expected value.
- If $\mu = E[X]$, the variance is defined as
    - $Var(X) = E[(X - \mu)^2] = E[X^2] - \mu^2$
- Variance is the second moment minus the square of the first moment.
- Variance, denoted by σ^2, satisfies the following property ($a, b \in \Re$):
    - $Var(aX + b) = a^2 Var(X)$

Standard Deviation

- $\sqrt{Var(X)}$ is called standard deviation and is denoted by σ.

- Standard deviation has the same unit as X and is easier to interpret than variance.

Covariance

- *Covariance* indicates the relationship between two random variables.
- If the occurrence of *X* makes *Y* more likely to occur, then the covariance is positive; it is negative if *X*'s occurrence makes *Y* less likely to happen and is 0 if there is no dependence.
  - $Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$
- where $\mu_X \equiv E[X]$ and $\mu_Y \equiv E[Y]$

Covariance

$$
\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathrm{Cov}(Y,X) \\
\mathrm{Cov}(X,X) &= \mathrm{Var}(X) \\
\mathrm{Cov}(X+Z,Y) &= \mathrm{Cov}(X,Y) + \mathrm{Cov}(Z,Y) \\
\mathrm{Cov}\left(\sum_i X_i, Y\right) &= \sum_i \mathrm{Cov}(X_i, Y)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y) \\
\mathrm{Var}\left(\sum_i X_i\right) &= \sum_i \mathrm{Var}(X_i) + \sum_i \sum_{j \neq i} \mathrm{Cov}(X_i, X_j)
\end{aligned}
$$

Covariance

If X and Y are independent, $E[XY] = E[X]E[Y] = \mu_X \mu_Y$ and $\mathrm{Cov}(X,Y) = 0$. Thus if X_i are independent

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i)$$

Correlation is a normalized, dimensionless quantity that is always between -1 and 1:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Random vectors

We can talk about multivariate distributions which give distributions of random vectors:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Expectation of a random vector

$$\mathrm{E}[X] = \begin{bmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \\ \vdots \\ \mathrm{E}[X_n] \end{bmatrix}$$

Random Vectors

The variance is generalized by the covariance matrix:

$$\mathbf{\Sigma} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \dots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \dots & \mathrm{Var}(X_n) \end{bmatrix}$$

Estimation of Parameters

- We make some assumptions about our problem by prescribing a parametric model (e.g. a distribution that describes how the data were generated), then we fit the parameters of the model to the data.
- How do we choose the values of the parameters?

Maximum likelihood estimation

Parametric Estimation

- $X = \{ x^t \}_t$ where $x^t \sim p(x)$
- Parametric estimation:

  Assume a form for $p(x \mid q)$ and estimate q, its sufficient statistics, using X

  e.g., $N(\mu, \sigma^2)$ where $q = \{ \mu, \sigma^2 \}$

Lecture notes by Ethem Alpaydın
Introduction to Machine Learning (Boğaziçi Üniversitesi)

Maximum Likelihood Estimation

- Likelihood of *q* given the sample X

  $l\ (\vartheta|X) = p\ (X\ |\vartheta) = \prod_t p\ (x^t|\vartheta)$

- Log likelihood

  $\mathcal{L}\ (\vartheta|X) = \log l\ (\vartheta|X) = \sum_t \log p\ (x^t|\vartheta)$

- Maximum likelihood estimator (MLE)

  $\vartheta^* = \text{argmax}_\vartheta\ \mathcal{L}\ (\vartheta|X)$

Lecture notes by Ethem Alpaydın
Introduction to Machine Learning (Boğaziçi Üniversitesi)

Maximum Likelihood Estimation

- For some distributions, it is possible to analytically solve for the maximum likelihood estimator.
- If $\mathcal{L}$ is differentiable, setting the derivatives to zero and trying to solve for θ is a good place to start.

Lecture notes by Ethem Alpaydın
Introduction to Machine Learning (Boğaziçi Üniversitesi)

Examples: Bernoulli/Multinomial

- Bernoulli: Two states, failure/success, x in {0,1}

  $P(x) = p_o^x (1 - p_o)^{(1-x)}$

  $$\mathcal{L}(p_o|\mathcal{X}) = \log \prod_t p_o^{x^t} (1 - p_o)^{(1-x^t)}$$

  MLE: $p_o = \sum_t x^t / N$

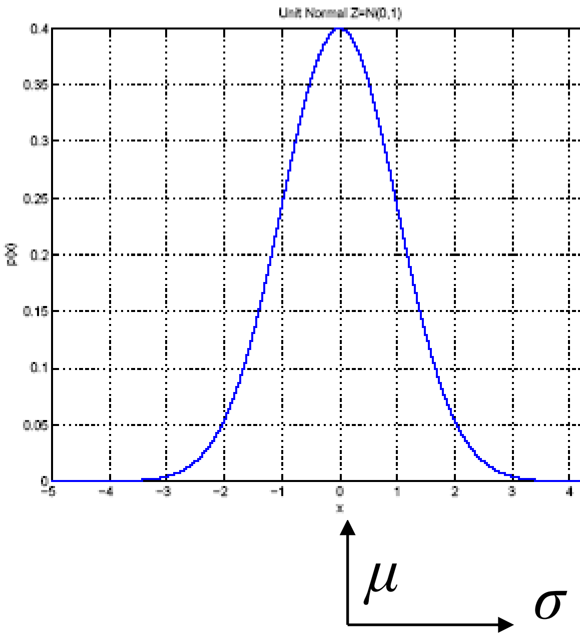
- Multinomial: K>2 states, x_i in {0,1}

  $P(x_1,x_2,\ldots,x_K) = \prod_i p_i^{x_i}$

  $$\mathcal{L}(p_1,p_2,\ldots,p_K|\mathcal{X}) = \log \prod_t \prod_i p_i^{x_i^t}$$

  MLE: $p_i = \sum_t x_i^t / N$

Gaussian (Normal) Distribution



- $p(x) = \mathcal{N}(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- MLE for μ and σ^2:

$$m = \frac{\sum_t x^t}{N}$$

$$s^2 = \frac{\sum_t (x^t - m)^2}{N}$$

Bias and Variance

Unknown parameter q
Estimator $d_i = d\ (X_i)$ on sample X_i

Bias: $b_q(d) = E\ [d] - q$
Variance: $E\ [(d - E\ [d])^2]$

Mean square error:
$r\ (d,q) = E\ [(d-q)^2]$
$= (E\ [d] - q)^2 + E\ [(d - E\ [d])^2]$
$=$ Bias^2 + Variance