CWM: AI and ML with python

Math review notes by Yangchen Pan

Linear Algebra

Matrix Product

Let A be an $m \times n$ matrix, B be an $n \times k$ matrix, a be an n-dimensional (column) vector, and b be an m-dimensional (column) vector.

Inner Product:
The inner product of vectors a and b is given by $a^T b = \sum_{i=1}^{n} a_i b_i$. It is important to note that this operation requires a and b to have the same dimensionality.

Outer Product:
The outer product of a and b is represented as $C = ab^T$, resulting in an $n \times m$ matrix. Here, the element in the i-th row and j-th column of C is $C_{ij} = a_i b_j$. This calculation does not require a and b to have equal dimensions.

Matrix-Vector Product:
The matrix-vector product when multiplying A^T by a yields a column vector r, with the i-th component defined as $r_i = c_i^T a$, where c_i is the i-th column of A (or the i-th row of A^T).

Matrix-Matrix Product:
The product of A and B is a matrix $C = AB$, where each element C_{ij} is computed as the inner product of the i-th row of A and the j-th column of B, i.e., $C_{ij} = A_{i,:} B_{:,j}$.

Expressing Matrix Product as Summation of Outer Products:
The matrix product can also be expressed as a summation of outer products: $AB = \sum_{i=1}^{n} A_{:,i} B_{i,:}$ where $A_{:,i}$ and $B_{i,:}$ are the i-th column and row of A and B respectively. This expression helps visualize how each element of C is the sum of products between corresponding columns of A and rows of B.

Common concepts/terms

1. Basis and Vector Space

Basis: In a vector space V , a set B of vectors is termed a basis if every element of V can be uniquely expressed as a finite linear combination of elements of B. The vectors in a basis are known as basis vectors.

Orthonormal Basis: In an orthonormal basis, all vectors are orthogonal to each other and each vector has a unit length.

Linear Combination: A linear combination of a set of vectors $\{v_1, v_2, \ldots, v_k\}$ in a vector space involves combining these vectors using scalar multiplication and vector addition. Specifically, a vector v is a linear combination of $\{v_1, v_2, \ldots, v_k\}$ if $v = c_1 v_1 + c_2 v_2 + \ldots + c_k v_k$, where $c_1, c_2, \ldots, c_k$ are scalars.

Norm: The norm of a vector v in a vector space, denoted as $\|v\|$, measures the “length” or “magnitude” of the vector. For a vector with components $(v_1, v_2, \ldots, v_n)$ in Euclidean space, the Euclidean norm (or L^2 norm) is defined as: $\|v\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}$ This norm is commonly used as it corresponds to the intuitive geometric length of a vector.

An extension is L^p norm, also known as the p-norm, is a generalization of the Euclidean norm (which is specifically the L^2 norm). It is defined for a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ in $\mathbb{R}^n$ and is used to measure the length (or "magnitude") of the vector in various ways, depending on the value of p, where p is a positive real number.

Definition:

The L^p norm of the vector $\mathbf{x}$ is defined as: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$

Properties:

1. **Non-negativity:** $\|\mathbf{x}\|_p \geq 0$ for all $\mathbf{x}$, and $\|\mathbf{x}\|_p = 0$ if and only if $\mathbf{x} = 0$.
2. **Scalar Multiplication:** For any scalar α, $\|\alpha\mathbf{x}\|_p = |\alpha| \cdot \|\mathbf{x}\|_p$.
3. **Triangle Inequality:** $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$.

2. Extension to Function Space

In function spaces, basis sets can consist of functions. For example, the set $\{x^n \mid n \in \mathbb{N}\}$ forms a basis in the space of polynomial functions.

Taylor Series: A practical application of function bases is the Taylor series expansion: $\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$ which expresses a function as an infinite sum of terms calculated from the values of its derivatives at a single point.

Other Examples:

- Fourier Series, where functions are expressed as sums of sines and cosines, which themselves form an orthonormal basis in the space of square-integrable functions over an interval.

3. Matrix Decomposition: Singular Value Decomposition (SVD)

A matrix X of size $n \times m$ can be decomposed via Singular Value Decomposition into $X = U\Sigma V^T$:

- U: an $n \times n$ orthogonal matrix.
- Σ: an $n \times m$ diagonal matrix with non-negative real numbers on the diagonal, known as singular values.
- V : an $m \times m$ orthogonal matrix.

Orthogonal Matrices: For any orthogonal matrix U, it holds that $UU^T = I = U^TU$, which implies $U^{-1} = U^T$. The same properties apply to matrix V .

Applications of SVD:

- Computing the pseudoinverse.
- Performing least squares minimization.
- Conducting low-rank approximations.

Expressing SVD as Summation of Outer Products: $X = U\Sigma V^T = \sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^T$ where u_i is the i-th column of U, σ_i is the i-th singular value, and v_i is the i-th column of V .

4. Linear Functions

In linear algebra, a linear function f is defined by the properties:

$$f(x + y) = f(x) + f(y)$$
$$f(ax) = af(x)$$

where x and y are vectors, and a is a scalar.

Example of a Linear Function: $f(x) = x^\top w$ This function is linear in both x and w.

Non-Linear Example: $f(x) = (x \circ x)^\top w$ Here, $\circ$ denotes the element-wise product. This function is not linear in x due to the element-wise squaring of x, which violates the linearity conditions.

Matrix calculus

A collection of calculations (relations, differentiations, etc.) of matrices: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Let x, b and w be n-dimensional column vectors; let A be an $m \times n$ matrix. Then Ax is an m-dimensional vector.

Computing the Gradient with Respect to x: $\nabla_x y$:

1. For $y = x^T w$, and $y = \frac{1}{2}x^T x$,

2. For $y = Ax$ (results in the Jacobian),
3. For $y = \frac{1}{2}\|Ax - b\|_2^2$,

Hessian Matrix

If the second-order derivatives of a function exist, then the Hessian matrix is defined as the matrix with elements in the ith row and jth column given by $\frac{\partial^2 y}{\partial x_i \partial x_j}$.

Probability

Common distributions

- Gaussian distribution function:

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

- Multivariate Gaussian: n-dimensional Gaussian random vector with constant diagonal covariance matrix

$$N(x; \mu, \sigma^2 I) = (2\pi)^{-n/2}\sigma^{-n} \ \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^\mathsf{T}(x-\mu)\right)$$

  Note that diagonal constant covariance means all entries in the Gaussian vector are independent and the variance is σ^2. This is exactly the same as the joint probability distribution of n Gaussian random variables.

A very frequently used property of Gaussian random vector: let A be a $n \times n$ constant matrix and b a n-dimensional constant vector, and $Y = AX + b$, then the $E[Y] = AE[X] + b$.

- Poisson distribution function (model count):

$$p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{where} \quad \lambda > 0 \text{ is the mean.}$$

- Bernoulli (model binary r.v.):

$$p(X = k) = p^k(1-p)^{1-k} \quad \text{for } k \in \{0, 1\}$$

- A very commonly seen assumption that allows to simplify calculation: independence assumption: $p(x_1, \ldots, x_n) = \prod_i p(x_i)$

Frequently used theorems

A random variable X is a function that assigns a real number to each outcome in a sample space. When we consider $X = x$, we are looking at the set of all outcomes in the sample space that map to the value x under the random variable X. This set of outcomes forms an event.

When we say $X = x$, we are identifying the event consisting of all sample points for which the random variable X takes the value x. This can be formally written as:

$$\{\omega \in \Omega : X(\omega) = x\}$$

where Ω is the sample space and ω represents an individual outcome in Ω.

Total Probability Theorem

$$P(X = x) = \sum_{i=1}^{n} P(X = x \mid Y = y_i) P(Y = y_i)$$

Conditional Probability

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

This means we are looking at the probability of the joint event $X = x$ and $Y = y$ occurring, given that $Y = y$ has occurred.

Bayes' Theorem

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

This theorem updates our belief about the probability of the event $X = x$ given that $Y = y$ has been observed.

Parameter estimation methods

Maximum likelihood estimation (MLE)

Maximum Likelihood Estimation (MLE) is a fundamental method used to estimate the parameters of a statistical model. The central idea is that we assume our observed data comes from a certain probability distribution with unknown parameters, and we want to estimate these parameters so that the observed data is most likely to occur.

Problem Setup

Consider a parameter estimation problem where we assume that a set of random variables $X_1, X_2, \ldots, X_n$ are drawn from a Gaussian distribution with an unknown mean μ and a known variance σ^2. The probability density function of a Gaussian distribution is given by:

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

Here, the semicolon (;) is used to denote that μ is a parameter of the distribution.

Given a set of observed values $x_1, x_2, \ldots, x_n$ of these random variables, our goal is to estimate the parameter μ.

Maximum Likelihood Estimation Approach

The high-level idea is to find the parameter μ that makes the observed data most probable.

Steps in MLE

1. **Construct the Likelihood Function:**

   The likelihood function represents the probability of observing the given data as a function of the parameter. Assuming the observations are independent, the likelihood function $L(\mu)$ can be written as:

   $$L(\mu) = p(x_1, x_2, \ldots, x_n; \mu) = \prod_{i=1}^{n} p(x_i; \mu)$$

2. **Maximize the Likelihood Function:**

   We aim to find the parameter μ that maximizes the likelihood function. Mathematically, this is expressed as:

   $$\hat{\mu} = \arg\max_{\mu} L(\mu) = \arg\max_{\mu} \prod_{i=1}^{n} p(x_i; \mu)$$

   The parameter $\hat{\mu}$ that maximizes this function is called the Maximum Likelihood Estimate (MLE).

3. **Log-Likelihood for Mathematical Convenience:**

   To simplify the calculations and avoid numerical issues like underflow, we often work with the log-likelihood function instead of the likelihood function. The log-likelihood function is:

   $$\log L(\mu) = \log\left(\prod_{i=1}^{n} p(x_i; \mu)\right) = \sum_{i=1}^{n} \log p(x_i; \mu)$$

   Hence, the MLE can also be found by maximizing the log-likelihood:

   $$\hat{\mu} = \arg\max_{\mu} \log L(\mu) = \arg\max_{\mu} \sum_{i=1}^{n} \log p(x_i; \mu)$$

   For convenience, this can also be expressed as the average log-likelihood:

   $$\hat{\mu} = \arg\max_{\mu} \frac{1}{n} \sum_{i=1}^{n} \log p(x_i; \mu)$$

In summary, MLE involves constructing the likelihood function based on the assumed distribution, and then finding the parameter that maximizes this likelihood. Using the log-likelihood simplifies the computation and helps avoid numerical issues.

Exercise 1: you observed five observations $u_1 = 1.30, u_2 = 2.12, u_3 = 2.40, u_4 = 0.98, u_5 = 1.43$ and you assume them from a uniform distribution $\sim U[0, \theta]$ where θ is the unknown parameter. Use MLE to estimate θ.

Exercise 2: derive the log-likelihood function of the previous Guassian distribution's mean (no need to solve the optimization).

Connection between KL Divergence and MLE*

The Kullback-Leibler (KL) divergence, $KL(P \parallel Q)$, measures how one probability distribution P diverges from a second, reference probability distribution Q. The definition of KL divergence is:

$$KL(P \parallel Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx = \mathbb{E}_{x \sim p}\left[\log\left(\frac{p(x)}{q(x)}\right)\right]$$

Let $p(x)$ be the underlying true distribution. To estimate the parameter in the distribution by minimizing the KL divergence, we write the distribution with the estimated parameter as $p(x; \mu')$.

The goal is to minimize the KL divergence:

$$\min_{\mu'} KL(p(x; \mu) \parallel p(x; \mu')) = \min_{\mu'} \mathbb{E}_{x \sim p}[\log p(x; \mu)] - \mathbb{E}_{x \sim p}[\log p(x; \mu')]$$

Since the first term $\mathbb{E}_{x \sim p}[\log p(x; \mu)]$ is a constant (independent of μ'), minimizing the KL divergence is equivalent to:

$$\max_{\mu'} \mathbb{E}_{x \sim p}[\log p(x; \mu')]$$

In practice, we approximate the expectation with the empirical average over n observed samples $x_1, x_2, \ldots, x_n$:

$$\max_{\mu'} \mathbb{E}_{x \sim p}[\log p(x; \mu')] \approx \max_{\mu'} \frac{1}{n} \sum_{i=1}^{n} \log p(x_i; \mu')$$

This shows that maximizing the likelihood (or equivalently, the log-likelihood) of the observed data is the same as minimizing the KL divergence between the true distribution and the estimated distribution. Thus, MLE can be viewed as a method to find the parameter μ' that makes the estimated distribution $p(x; \mu')$ as close as possible to the true distribution $p(x)$ in terms of KL divergence.

Maximum A Posteriori estimation (MAP) *

We assume a prior distribution on the parameter, say $g(\mu)$. After observing the data, we update prior belief by maximizing the posterior distribution $p(\mu \mid x)$.

Now, assume that a prior distribution $g(\mu)$ over μ exists. This allows us to treat μ as a random variable, as in Bayesian statistics. We can calculate the posterior distribution of μ using Bayes' theorem:

$$p(\mu \mid x) = \frac{p(x \mid \mu) g(\mu)}{p(x)}$$

Then the MAP estimate is:

$$\begin{aligned}
\hat{\mu_{\text{MAP}}} &= \arg\max_{\mu} \; p(\mu \mid x) \\
&= \arg\max_{\mu} \; \frac{p(x \mid \mu)\, g(\mu)}{\displaystyle\int p(x \mid \mu')\, g(\mu')\, d\mu'} \\
&= \arg\max_{\mu} \; p(x \mid \mu)\, g(\mu).
\end{aligned}$$

Bayes theorem: $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$ where $P(B) = \sum_{A_i} P(B|A_i)P(A_i)$ if A_i s are partition of the set A.

Summary: Two Views of Statistical Estimation

1. Frequentist View: Maximum Likelihood Estimation (MLE)

In the frequentist approach, we assume that the unknown parameter is a **constant**. Our goal is to estimate this fixed parameter using statistical methods. This is the basis of **frequentist statistics**, which views parameters as fixed quantities that need to be estimated from the data.

- **MLE Setup:** In this setup, we assume that the unknown parameter θ is fixed but unknown. We then use methods like Maximum Likelihood Estimation (MLE) to estimate it.
- **Frequentist Assumption:** The world is considered to be fixed and constant, but unknown, and we need to use statistical techniques to estimate these fixed quantities.

2. Bayesian View: Bayesian Inference

In the Bayesian approach, the unknown parameter is considered a **random variable**. Given a training set S, we compute the posterior distribution of the parameter based on observed data and prior beliefs.

- **Bayesian Assumption:** The unknown parameter θ is treated as a random variable. This allows us to incorporate prior knowledge and update our beliefs with observed data.
- **Posterior Distribution:** Using Bayes' theorem, we calculate the posterior distribution of θ:

  $$p(\theta \mid S) = \frac{p(S \mid \theta)g(\theta)}{p(S)}$$

- **Predictive Distribution:** When given a new testing data point x, we compute the posterior distribution of the class label y using the posterior distribution on θ:

  $$p(y \mid x, S) = \int_\theta p(y \mid x, \theta) p(\theta \mid S)\, d\theta$$

  This integral can be challenging to compute directly.

- **MAP Method:** One practical approach to avoid the complexity of integration is to use the Maximum A Posteriori (MAP) estimate. This method finds a point estimate of θ that maximizes the posterior distribution, providing a simplified yet effective estimation.