3.1 l-norms and Distance Metrics

3.1.1 l-norms

$\mathbf{x}$ is a column vector in $\mathbb{R}^N$ space.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_N \end{bmatrix}$$

l_0-norm

$$\|\mathbf{x}\|_0 \equiv \sum_{i=1}^{N} |x_i|^0$$

$$= |x_1|^0 + |x_2|^0 + \cdots + |x_N|^0$$

ℹ️ l_0-norm is the *number of non-zero elements* in vector X.

- Defined that $0^0 = 0$.

- Application:
  - $\|\mathbf{x}\|_0$ is very small $\iff$ The vector $\mathbf{x}$ is very sparse/shallow.
  - Minimize $\|\mathbf{x} - \mathbf{y}\|_0 \iff$ Minimize the difference between $\mathbf{x}$ and $\mathbf{y}$.

l_1-norm (Taxicab Norm / Manhattan Norm)

$$\|\mathbf{x}\|_1 \equiv \sum_{i}^{N} |x_i|$$

$$= |x_1| + |x_2| + \cdots + |x_N|$$

ℹ️ l_1-norm is the *sum of elements' absolute values* in vector X.

- Application:
  - Minimize $\|\mathbf{x}\|_1 \iff$ Minimize total value of non-zero element sums. Similar results as minimize $\|\mathbf{x}\|_0$.

l_2-norm (Euclidean Norm)

$$\|\mathbf{x}\|_2 \equiv (\sum_{i=1}^{N} |x_i|^2)^{\frac{1}{2}}$$

$$= \sqrt{x_1^2 + x_2^2 + \cdots + x_N^2}$$

- l_2-norm can be expressed as matrix format $\|\mathbf{x}\|_2 \equiv \sqrt{\mathbf{x}^\top \mathbf{x}}$
- Application:
  - Minimize $\|X\|_2 \iff$ Make matrix more sparse.

l_∞-norm (Maximum Norm)

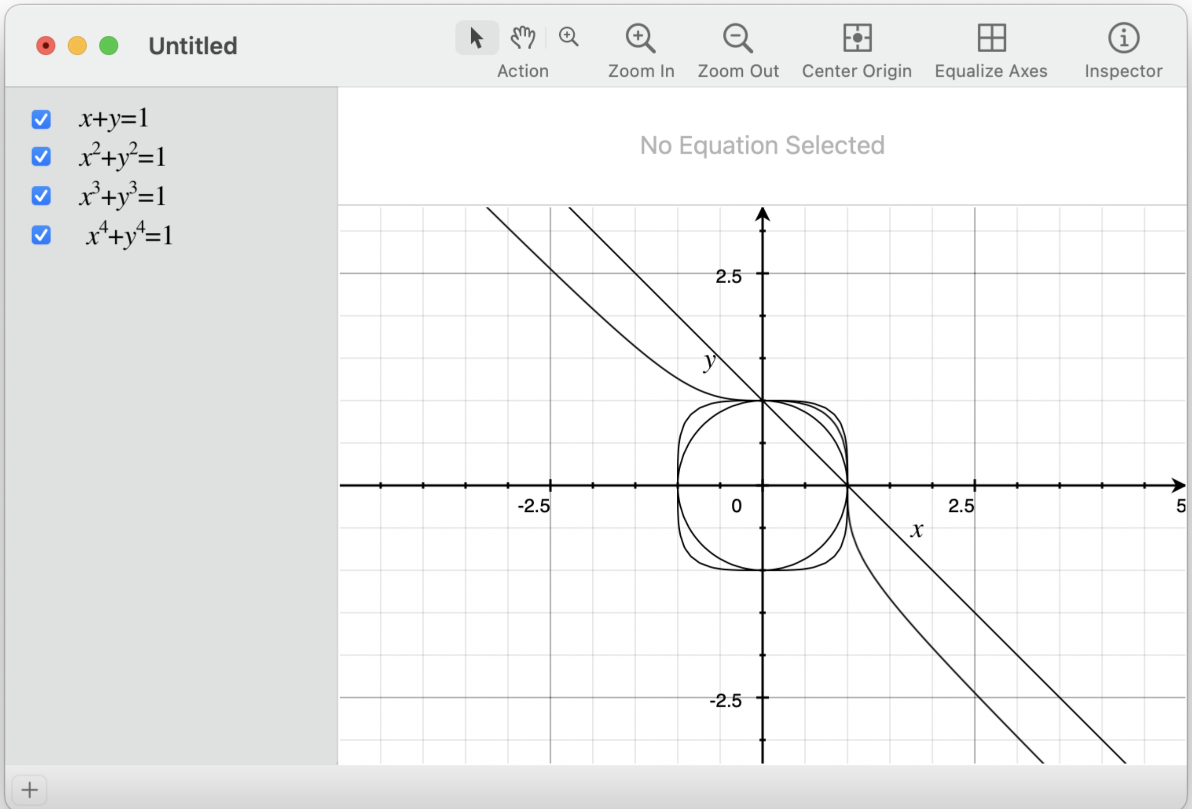
$$\|\mathbf{x}\|_\infty \equiv \max(|x_1|, |x_2|, \cdots, |x_N|)$$

- l_∞-norm takes the *maximum of absolute values of elements* in vector $\mathbf{x}$.
- l_∞-norm is also called
  - Maximum norm

l_p-norm

$$\|X\|_p \equiv \left(\sum_{i=1}^{N} |x_i|^p\right)^{\frac{1}{p}} = (|x_1|^2 + |x_2|^2 + \cdots + |x_N|^p)^{\frac{1}{p}}$$

- l_p-norm is a general form of l-norm, where $p \geq 0$.
  - $p = 0$, l_0-norm,
  - $p = 1$, l_1-norm,
  - $p = 2$, l_2-norm,
  - ...,
  - $p \to \infty$, l_∞-norm.



3.1.2 Distance Metrics

Euclidean Distance

Given

- Two datasets $\mathbf{x}, \mathbf{y}$:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_N \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} \in \mathbb{R}^N$$

Do

Euclidean Distance:

$$
\begin{aligned}
d_E(\mathbf{x}, \mathbf{y}) &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_N - y_N)^2} \\
&= \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \\
&= \sqrt{(\mathbf{x} - \mathbf{y})^\top (\mathbf{x} - \mathbf{y})}
\end{aligned}
$$

- The straight-line distance between X and Y.
- Also called L_2 distance.

Mahalanobis Distance

Given

- An observation.

$$
\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_N \end{bmatrix}
$$

- A set of observations with
  - mean $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_N \end{bmatrix}$
  - Covariance matrix Σ

Do

Mahalanobis Distance

$$
d_M(\mathbf{x}, \mu) = \sqrt{(\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu)}
$$

- It is a measure of distance between:
  - a point, and
  - a distribution
- It reverts to Euclidean Distance when $\Sigma = I$.

3.2 Parameter Estimation

Recall the Bayes Formula:

$$
P(\omega_j | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_j) P(\omega_j)}{P(\mathbf{x})}
$$

All we have initially are the training samples.

- We don't directly "know" the prior & posterior probabilities.
- Therefore, we need to retrieve prior probability $P(\omega_j)$ and posterior probability $P(X|\omega_j)$ from training samples.

Collect training samples $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ distributed according to the unknown $P(\mathbf{x}|\omega_j)$.

Assumed that these samples are Independent and identically distributed (i.i.d.).

- Independent: $\mathbf{x}_i$ and $\mathbf{x}_j$ does not influence each other.
- Identical: $\mathbf{x}_i \neq \mathbf{x}_j,\ \forall i \neq j$.

Our next goal is to estimate μ_j and Σ_j, hyper parameters of the posterior distribution $P(\mathbf{x}|\omega_j)$.

- Parametric Form
  - Maximum Likelihood Estimation (MLE)
  - Bayesian Estimation (BE)
- Nonparametric Form

3.3 Maximum Likelihood Estimation

3.3.1 Find the best θ : Log-Likelihood.

Given

- The set of i.i.d. training Examples:

$$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$$

- where:
  - $\forall k = 1, 2, \cdots, N,\ \mathbf{x}_k \sim P(\mathbf{x}|\theta)$
- θ are the parameters to be estimated.

  Do

  We derive the objective function:

$$p(\mathcal{X}|\theta) \equiv p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N|\theta)$$

$$= \prod_{k=1}^{N} p(\mathbf{x}_k|\theta)$$

  $p(X|\theta)$ is the Likelihood of θ with respect to X.

To find a best θ, we derive a Maximum Likelihood estimation:

$$\hat{\theta}_{ML} = \text{argmax}_\theta \, p(\mathcal{X}|\theta)$$

$$= \text{argmax}_\theta \prod_{k=1}^{N} p(\mathbf{x}_k|\theta)$$

That is, we want to find the θ that gives the Maximum Likelihood on $\mathcal{X}$.

Log-likelihood

For optimization purposes, we derive a log-likelihood function that preserves the monotonicity of the original MLE.

$$L(\theta) = \ln p(\mathcal{X}|\theta)$$

$$= \ln \prod_{k=1}^{N} p(\mathbf{x}_k|\theta)$$

$$= \sum_{k=1}^{N} \ln p(\mathbf{x}_k|\theta)$$

As the monotonicity is preserved, we would derive that

$$\hat{\theta}_{ML} = \text{argmax}_\theta \, p(\mathcal{X}|\theta)$$

$$= \text{argmax}_\theta \, L(\theta)$$

$$= \text{argmax}_\theta \sum_{k=1}^{N} \ln \; p(\mathbf{x}_k|\theta)$$

Equivalently, we find the θ that gives the *maximum* $L(\theta)$ now.

To find the θ that maximizes $L(\theta)$, we find:

$$\hat{\theta}_{ML} : \; \frac{\partial L(\theta)}{\partial \theta} = 0$$

That is,

$$\hat{\theta}_{ML} : \; \sum_{k=1}^{N} \frac{\partial \left[\ln p(x_k|\theta)\right]}{\partial \theta} = 0$$

3.3.2 μ unknown, Σ known; $\theta = \{\mu\}$.

Univariate & Multivariate Case ($x \in \mathbb{R}^{N^+}$)

The distribution:

$$p(\mathbf{x}_k|\mu) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}_k-\mu)^\top\Sigma^{-1}(\mathbf{x}_k-\mu)}$$

The log likelihood:

$$\ln p(\mathbf{x}_k|\mu) = -\frac{1}{2}\ln\Big((2\pi)^d|\Sigma|\Big) - \frac{1}{2}(\mathbf{x}_k-\mu)^\top\Sigma^{-1}(\mathbf{x}_k-\mu)$$

$$= -\frac{1}{2}\Big((2\pi)^d|\Sigma|\Big) - \frac{1}{2}(\mathbf{x}_k^\top-\mu^\top)\Sigma^{-1}(\mathbf{x}_k-\mu)$$

$$= -\frac{1}{2}\Big((2\pi)^d|\Sigma|\Big) - \frac{1}{2}(\mathbf{x}_k^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})(\mathbf{x}_k-\mu)$$

$$= -\frac{1}{2}\Big((2\pi)^d|\Sigma|\Big) - \frac{1}{2}(\mathbf{x}_k^\top\Sigma^{-1}\mathbf{x}_k-\mathbf{x}_k^\top\Sigma^{-1}\mu-\mu^\top\Sigma^{-1}\mathbf{x}_k+\mu^\top\Sigma^{-1}\mu)$$

$$= -\frac{1}{2}\Big((2\pi)^d|\Sigma|\Big) - \frac{1}{2}\mathbf{x}_k^\top\Sigma^{-1}\mathbf{x}_k+\mu^\top\Sigma^{-1}\mathbf{x}_k-\frac{1}{2}\mu^\top\Sigma^{-1}\mu$$

The constant terms are:

- $-\frac{1}{2}\Big((2\pi)^d|\Sigma|\Big)$
- $-\frac{1}{2}\mathbf{x}_k^\top\Sigma^{-1}\mathbf{x}_k$, since $\mathbf{x}_k$ is pre-defined.

Therefore:

$$\frac{\partial}{\partial\mu}\ln p(\mathbf{x}_k|\mu) = \frac{\partial}{\partial\mu}(\mu^\top\Sigma^{-1}\mathbf{x}_k-\frac{1}{2}\mu^\top\Sigma^{-1}\mu)$$

$$= \Sigma^{-1}\mathbf{x}_k-\Sigma^{-1}\mu$$

$$= \Sigma^{-1}(\mathbf{x}_k-\mu)$$

As we required

$$\frac{\partial}{\partial \mu} L(\mu) = 0$$

$$\implies \sum_{k=1}^{N} \frac{\partial}{\partial \mu} \ln p(\mathbf{x}_k | \mu) = 0$$

$$\implies \sum_{k=1}^{N} \Sigma^{-1} (\mathbf{x}_k - \mu) = 0$$

$$\implies \Sigma^{-1} \sum_{k=1}^{N} (\mathbf{x}_k - \mu) = 0$$

$$\implies \left(\sum_{k=1}^{N} \mathbf{x}_k \right) - N\mu = 0$$

$$\implies N\mu = \sum_{k=1}^{N} \mathbf{x}_k$$

$$\implies \mu = \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k$$

★ That is, the μ that produces the maximum likelihood over the dataset is:

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_k$$

3.3.3 μ unknown, Σ unknown; $\theta = \{\mu, \Sigma\}$

Univariate Case ($x \in \mathbb{R}$)

$$p(x_k | \theta) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x_k - \mu)^2}{2\sigma^2}}$$

where:

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$$

The log likelihood:

$$\ln p(x_k|\theta) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_k - \mu)^2$$

$$= -\frac{1}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

Therefore:

$$\frac{\partial}{\partial\theta}\ln p(x_k|\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial\theta_1} \\ \frac{\partial L(\theta)}{\partial\theta_2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{(x_k - \theta_1)}{\theta_2} \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

Again, to find the θ that minimizes the MLE, we let

$$\frac{\partial}{\partial\theta}L(\theta) = 0$$

$$\Longrightarrow \sum_{k=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_k|\theta) = 0$$

$$\Longrightarrow \begin{bmatrix} \sum_{k=1}^{N}\frac{x_k - \theta_1}{\theta_2} \\ \sum_{k=1}^{N}\left(-\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2}\right) \end{bmatrix} = 0$$

For the first term:

$$\sum_{k=1}^{N}\frac{x_k - \theta_1}{\theta_2} = 0$$

$$\Longrightarrow \sum_{k=1}^{N}(x_k - \theta_1) = 0$$

$$\Longrightarrow (\sum_{k=1}^{N}x_k) - N\theta_1 = 0$$

$$\Longrightarrow \theta_1 = \frac{1}{N}\sum_{k=1}^{N}x_k$$

★ That is, the optimal μ is:

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

For the second term:

$$\sum_{k=1}^{N} -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} = 0$$

$$\implies \sum_{k=1}^{N} -\theta_2 + (x_k-\theta_1)^2 = 0$$

$$\implies -N\theta_2 + \sum_{k=1}^{N}(x_k-\theta_1)^2 = 0$$

$$\implies \theta_2 = \frac{1}{N}\sum_{k=1}^{N}(x_k-\theta_1)^2$$

★ That is, the optimal σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k-\hat{\mu})^2$$

Multivariate Case ($x \in \mathbb{R}^D, D > 1$)

$$p(\mathbf{x}_k|\theta) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}_k-\mu)^\top \Sigma^{-1}(\mathbf{x}_k-\mu)}$$

where:

$$\theta = \begin{bmatrix}\mu \\ \Sigma\end{bmatrix} = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix}$$

★ Similarly, we have the optimized parameters as:

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k-\hat{\mu})(\mathbf{x}_k-\hat{\mu})^\top$$

3.4 Bayesian Estimation

3.4.0 Difference between BE and MLE

- In ML estimation, θ was considered a parameter with a fixed value.
- In Bayesian estimation however, θ is considered an unknown random vector.
    - which is described by a P.D.F $p(\theta)$.

3.4.1 Find the best θ : Bayes Formula

Given

- The set of i.i.d. training examples:

$$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$$

- where:
    - $\forall k = 1, 2, \cdots, N,\ \mathbf{x}_k \sim P(\mathbf{x}|\theta)$
- θ are the parameters to be estimated.

Do

As θ is regarded to be random, we compute the maximum of $p(\theta|\mathcal{X})$. From Bayes formula, we know that:

$$p(\theta|\mathcal{X}) = \frac{p(\mathcal{X}|\theta) \cdot P(\theta)}{P(\mathcal{X})}$$

We find the with the best Maximum Aposterior Probability.

$$\hat{\theta}_{MAP} = \text{argmax}_\theta\, p(\theta|\mathcal{X})$$

$$= \text{argmax}_\theta\, p(\mathcal{X}|\theta) \cdot P(\theta)$$

Similarly, find the max:

$$\frac{\partial}{\partial\theta} \ln p(\theta|\mathcal{X}) = 0$$

$$\implies \frac{\partial}{\partial\theta} \ln\Big(p(\mathcal{X}|\theta) \cdot P(\theta)\Big) = 0$$

3.4.2 μ unknown, σ known

Univariate case

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

where μ conforms a normal distribution:

$$\mu \sim N(\mu_0, \sigma_0) : \; p(\mu) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

We could therefore know that:

$$\ln p(x_k|\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_k - \mu)^2$$

$$\ln P(\mu) = -\frac{1}{2}\ln(2\pi\sigma_0^2) - \frac{1}{2\sigma^2}(\mu - \mu_0)^2$$

Therefore,

$$\frac{\partial}{\partial\mu}\ln p(x_k|\mu) = \frac{(x_k - \mu)}{\sigma^2}$$

$$\frac{\partial}{\partial\mu}\ln P(\mu) = \frac{(\mu - \mu_0)}{\sigma_0^2}$$

Therefore, the optimal μ could be obtained by:

$$\frac{\partial}{\partial \mu}\ln\Big(p(\mathcal{X}|\mu)\cdot P(\mu)\Big) = 0$$

$$\implies \frac{\partial}{\partial \mu}\ln\Big[\prod_{k=1}^{N} p(\mathbf{x}_k|\mu)\cdot P(\mu)\Big] = 0$$

$$\implies \frac{\partial}{\partial \mu}\Big[\sum_{k=1}^{N}\ln p(\mathbf{x}_k|\mu)\Big] + \frac{\partial}{\partial \mu}\ln P(\mu) = 0$$

$$\implies \Big[\sum_{k=1}^{N}\frac{\partial}{\partial \mu}\ln p(\mathbf{x}_k|\mu)\Big] + \frac{\partial}{\partial \mu}\ln P(\mu) = 0$$

$$\implies (\sum_{k=1}^{N}\frac{x_k-\mu}{\sigma^2}) - (\frac{\mu-\mu_0}{\sigma_0^2}) = 0$$

$$\implies \frac{1}{\sigma^2}(\sum_{k=1}^{N}x_k) - \frac{N}{\sigma^2}\mu - \frac{1}{\sigma_0^2}\mu + \frac{1}{\sigma_0^2}\mu_0 = 0$$

$$\implies \sigma_0^2(\sum_{k=1}^{N}x_k) - (\sigma_0^2 N + \sigma^2)\mu + \sigma^2\mu_0 = 0$$

$$\implies \sigma_0^2(\sum_{k=1}^{N}x_k) + \sigma^2\mu_0 = (\sigma_0^2 N + \sigma^2)\mu$$

$$\implies \mu = \frac{\sigma_0^2(\sum_{k=1}^{N}x_k) + \sigma^2\mu_0}{\sigma_0^2 N + \sigma^2}$$

$$\implies \mu = \frac{\frac{\sigma_0^2}{\sigma_2}\sum_{k=0}^{N}x_k + \mu_0}{\frac{\sigma_0^2}{\sigma^2}N + 1}$$

★ That is, the optimal μ by Bayesian Estimation is:

$$\hat{\mu}_{BE} = \frac{\frac{\sigma_0^2}{\sigma_2}\sum_{k=0}^{N}x_k + \mu_0}{\frac{\sigma_0^2}{\sigma^2}N + 1}$$