1 Definitions For Real!

1.1 Regularization

1. ***L2-norm***: Also known as the Euclidean norm, is a measure of the magnitude of a vector in Euclidean space. It is calculated as:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{N} x_i^2} \tag{1}$$

where N is the dimensionality of the vector $\mathbf{x}$.

2. ***L1-norm***: a measure of the absolute magnitude of a vector. It is calculated as:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{N} |x_i| \tag{2}$$

again, where N is the dimensionality of the vector $\mathbf{x}$.

3. ***L0-norm***: also known as the "counting norm", is a measure of the number of non-zero elements in a vector. It is calculated as:

$$\|x\|_0 = \sum_{i=1}^{N} \mathbb{I}(x_i \neq 0) \tag{3}$$

where $\mathbb{I}(x_i \neq 0)$ is the indicator function that is equal to 1 if $x_i \neq 0$ and 0 otherwise.

1.2 Probability and Statistics Review

You should be familiar with most of this stuff.

(a) ***Random Variables***: In this class, we will be using random variable notation. A random variable is a mapping of events to values, and then the associated pmf (or pdf) maps those values to probabilities (or densities). For example, if Z takes on the value 1 when the roll of a six-sided fair dice is even, the probability of rolling an even number will be denoted as the following:

$$P(Z = 1) = \frac{1}{2}$$

(b) ***Conditional probability***: Probability distribution of a variable given another variable takes on a certain value

$$P(A = a \mid B = b) = \frac{P(A = a, B = b)}{P(B = b)}$$

(c) ***Marginalization (Sum Rule)***: Probability distribution of a single variable in a joint distribution

$$P(A = a) = \sum_{b \in \{\text{all values of } Y\}} P(A = a, B = b)$$

(d) ***Law of Total Probability***: Using conditional probability we can rewrite the above expression for marginalization as:

$$P(A = a) = \sum_{b} P(A = a \mid B = b) P(B = b)$$

(e) ***Law of Total Expectation***: For two given random variables A, B:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$$

If Y is discrete, this becomes:

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y]P(Y = y)$$

(f) ***Continuous vs discrete random variables***: Discrete random variables can only take a countable number of values (e.g., values we can roll in a dice) while continuous can take on infinitely many values. Our definitions for marginalization and the law of total probability assumed Y was a discrete random variable. If Y is not:

$$P(A = a) = \int_b P(A = a, B = b) * db$$

(g) ***Independence***: If two R.V.s are considered to be independent Then the following hold true $\forall a, b$: $P(A = a, B = b) = P(A = a)P(B = b)$ and $P(A = a \mid B = b) = P(A = a)$ and $P(B = b \mid A = a) = P(B = b)$.

(h) ***Conditional Independence:*** If two random variables A and B are conditionally independent given C then they are independent after conditioning on C

$$P(A, B \mid C) = P(B \mid A, C) * P(A \mid C) = P(B \mid C) * P(A \mid C)$$

(i) ***Variance:*** It's a measure of spread for a distribution of a random variable that determines the degree to which the values of a random variable differ from the expected value.

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

$Var(X) = \sigma^2$ where σ is the standard deviation of X

(j) ***Gaussian Distribution:*** Denoted as $X \sim \mathcal{N}(\mu, \sigma)$, the probability density function is given by: $p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2\sigma^2}(x - \mu)^2)$

(k) ***Independent and identically distributed (i.i.d.)***: A collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent i.e. each variable has the same chance of occurring as the others, and none of them have an influence on one another. This is a popular and important concept in statistics. A lot of the models and algorithms assume this property about their data.

2 Regularization

Think back to linear regression. We would calculate our objective function that we want to minimize as follows: $J(\boldsymbol{\theta}) = \frac{1}{N}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$.

Consider we add L2 regularization to this objective function.
This would give us: $J(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$. Determine the closed-form solution of this objective function.

We can expand our objective function as follows:

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum_{i=1}^{N}(Y_i - X_i\boldsymbol{\theta})^2 + \lambda\|\boldsymbol{\theta}\|_2^2 \\ &= (\mathbf{y} - X\boldsymbol{\theta})^\top(\mathbf{y} - X\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_2^2 \\ &= (\mathbf{y} - X\boldsymbol{\theta})^\top(\mathbf{y} - X\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta} \end{aligned}$$

Because we want to optimize this, we take the derivative of $J(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and set it to 0, to get our closed-form solution

$$\begin{aligned} \nabla J(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}(\mathbf{y} - X\boldsymbol{\theta})^\top(\mathbf{y} - X\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta} \\ &= \frac{\partial}{\partial\boldsymbol{\theta}}(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top X\boldsymbol{\theta} - \boldsymbol{\theta}^\top X^\top\mathbf{y} + \boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta} + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}) \\ &= \frac{\partial}{\partial\boldsymbol{\theta}}(\mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\theta}^\top X^\top\mathbf{y} + \boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta} + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}) \\ &= 0 - 2X^\top\mathbf{y} + 2X^\top X\boldsymbol{\theta} + \frac{\partial}{\partial\boldsymbol{\theta}}(\lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}) \\ &= -2X^\top(\mathbf{y} - X\boldsymbol{\theta}) + 2\lambda\boldsymbol{\theta} \end{aligned}$$

$$\begin{aligned} -2X^\top(\mathbf{y} - X\boldsymbol{\theta}) + 2\lambda\boldsymbol{\theta} &= 0 \\ -X^\top\mathbf{y} + X^\top X\boldsymbol{\theta} + \lambda\boldsymbol{\theta} &= 0 \\ X^\top X\boldsymbol{\theta} + \lambda\boldsymbol{\theta} &= X^\top\mathbf{y} \\ \cancel{(X^\top X + \lambda)}\boldsymbol{\theta} &= X^\top\mathbf{y} \\ (X^\top X + \lambda I)\boldsymbol{\theta} &= X^\top\mathbf{y} \\ \boldsymbol{\theta} &= (X^\top X + \lambda I)^{-1}X^\top\mathbf{y} \end{aligned}$$

Hence the closed form solution of our objective function with L2 regularization included is $\boldsymbol{\theta} = (X^\top X + \lambda I)^{-1}X^\top\mathbf{y}$

3 Multivariate Gaussian Distribution

3.1 Covariance Matrix

You may be familiar with Gaussian distributions for scalar random variables, i.e., one-dimensional, but we can also have N-dimensional random vector, $X = [X_1, X_2, \ldots, X_N]^T$. For this, we use multivariate Gaussian distribution. Notice we cannot use the normal variance, σ^2, anymore, so we use a covariance matrix. The covariance matrix, C, has the dimensions $N \times N$ and $C_{i,j} = cov[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$

What do the diagonal elements represent?

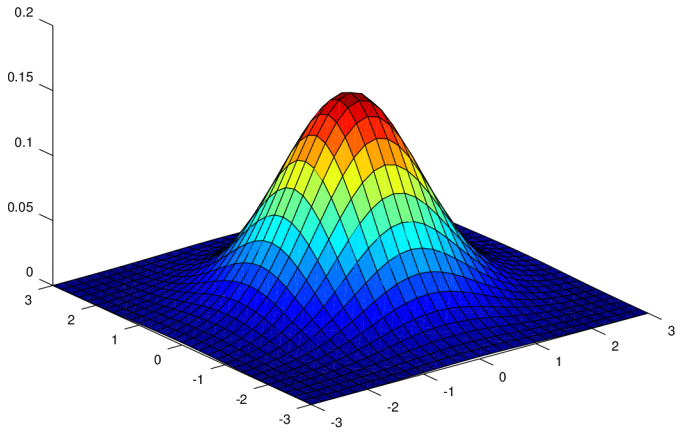
$C_{i,i} = Cov(X_i, X_i) = \mathbb{E}[X_i X_i] - \mathbb{E}[X_i]\mathbb{E}[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = Var(X_i)$
So, they represent the variance of the respective element.

What are some special properties of the covariance matrix?

It's always both symmetric (i.e., $C_{i,j} = C_{j,i}\ \forall i, j$) and positive semi-definite (i.e., $\mathbf{z}^\top C\mathbf{z} \geq 0 \forall \mathbf{z} \in \mathbb{R}^N$).

3.2 pdf

It would look something like this:



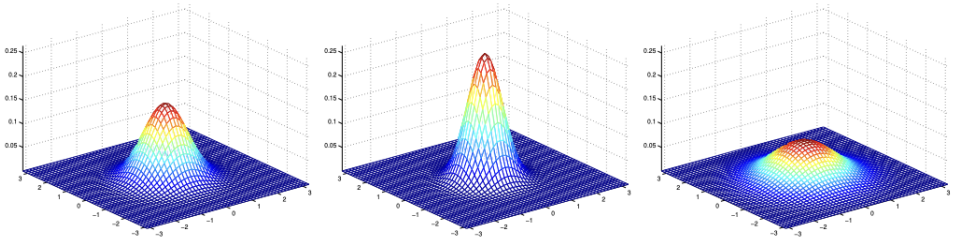
where the probability density function (pdf) is given by:

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}))$$

$\boldsymbol{\mu} \in \mathbb{R}^N$ is the mean and $\Sigma \in \mathbb{R}^{N \times N}$ is the covariance matrix. $|\Sigma|$ refers to the determinant of Σ.

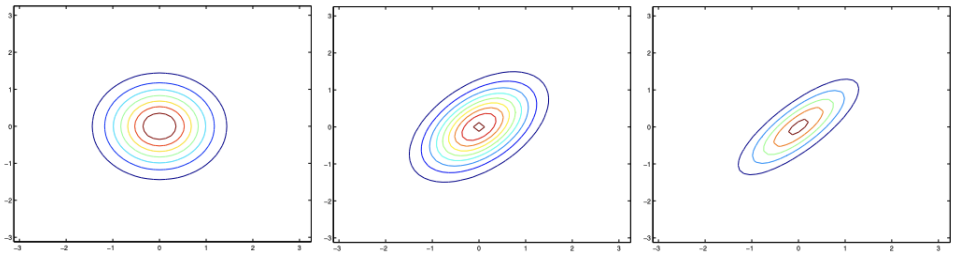
The following $\boldsymbol{\mu}$ and Σ values correspond to these multivariate Gaussian examples:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$



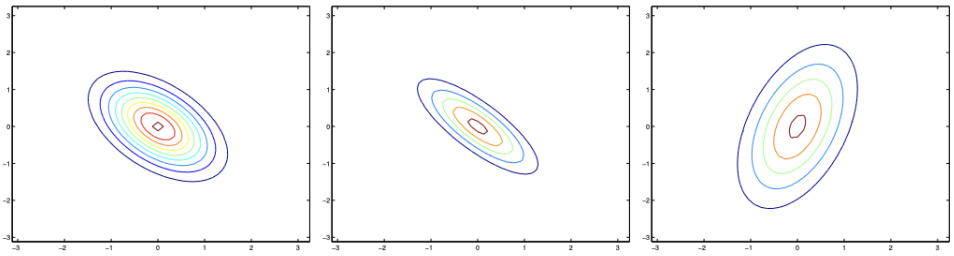
We can look at the contour lines to analyze Σ.

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}:$$



Similarly,

$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 3 & 0.8 \\ 0.8 & 1 \end{bmatrix}:$$



3.3 Multivariate Gaussian: What affects the "spread" of the distribution across the diagonal?

The covariance between variables X_1 and X_2 affects the spread of the distribution across the diagonal. A distribution that is more concentrated along the diagonal would have larger covariance between the two variables A negative covariance between two variables would reverse the direction of the diagonal.

4 Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model by maximizing the likelihood function.

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta; \mathcal{D}))$$

$\mathcal{D}$ is our data sample that we observe, $\mathcal{D} = \{y^{(i)}\}_{i=1}^{N}$.

$\mathcal{L}(\theta)$ is the likelihood of observing our sampled data points for a specific parameter θ, i.e., it is the joint density of the observed data sample given the parameters, $p(y^{(1)}, y^{(2)}, \ldots, y^{(N)} \mid \theta)$. $\mathcal{L}$ is a real-valued function of θ so we want to pick the values for our parameters θ that maximizes the likelihood function. With more data points (larger N), we can arrive at parameters that are closer to their true values.

Oftentimes, we take the log of the $\mathcal{L}$ because maximizing the log function is easier. And, maximizing the log function would also maximize the likelihood function.

Very Important!!!: We assume that the data is i.i.d. here

4.1 Bernoulli

Suppose we are flipping an unbiased coin. We can consider each flip to be a Bernoulli(ϕ) random variable, Y, where $Y = 1$ represents the coin coming up heads and Y=0 represents tails, and $P(Y = 1) = \phi$ and $P(Y = 0) = 1 - \phi$

Suppose you flipped heads N_H times and tails N_T times. Derive the log-likelihood function.

$$\begin{aligned}
\mathcal{L}(\phi) &= \prod_{i=1}^{N} \phi^{\mathbb{I}(y^{(i)}=1)}(1-\phi)^{\mathbb{I}(y^{(i)}=0)} \\
\ell(\phi) &= \log \prod_{i=1}^{N} \phi^{\mathbb{I}(y^{(i)}=1)}(1-\phi)^{\mathbb{I}(y^{(i)}=0)} \\
&= \sum_{i=1}^{N} \left[\log \phi^{\mathbb{I}(y^{(i)}=1)} + \log(1-\phi)^{\mathbb{I}(y^{(i)}=0)}\right] \\
&= \sum_{i=1}^{N} \left[\mathbb{I}(y^{(i)}=1)\log(\phi) + \mathbb{I}(y^{(i)}=0)\log(1-\phi)\right] \\
&= \log(\phi)\sum_{i=1}^{N}\mathbb{I}(y^{(i)}=1) + \log(1-\phi)\sum_{i=1}^{N}\mathbb{I}(y^{(i)}=0) \\
&= N_H \log(\phi) + N_T \log(1-\phi)
\end{aligned}$$

How can you represent the MLE of the probability of flipping heads, $\hat{\phi}_{MLE}$

Now that we have the log-likelihood function, our $\hat{\phi}_{MLE}$ is the value that maximizes the log-likelihood. Thus, we can take the derivative of our function and set it to 0.

$$\frac{d\ell}{d\phi} = N_H \frac{1}{\phi} - N_T \frac{1}{1-\phi} = \frac{(1-\phi)N_H - \phi N_T}{\phi(1-\phi)}$$

$$\frac{(1-\phi)N_H - \phi N_T}{\phi(1-\phi)} = 0$$

$$(1-\phi)N_H - \phi N_T = 0$$

$$N_H - \phi(N_H + N_T) = 0$$

$$\phi = \frac{N_H}{N_H + N_T}$$

Therefore, $\hat{\phi}_{MLE} = \frac{N_H}{N_H+N_T}$

4.2 Gaussian

Suppose we observe N samples from a Gaussian distribution, $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$, where $x^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma)$. Recall, that in a Gaussian distribution, the pdf is of the form:

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

Derive the log-likelihood function

$$\begin{aligned}
\mathcal{L}(x \mid \mu, \sigma) &= \prod_{i=1}^N p(x \mid \mu, \sigma) \\
&= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\right) \\
\ell(x \mid \mu, \sigma) &= \log \prod_{i=1}^N p(x \mid \mu, \sigma) \\
&= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^N -\log(\sqrt{2\pi}) - \log(\sigma) - \frac{(x^{(i)}-\mu)^2}{2\sigma^2}
\end{aligned}$$

Use the log-likelihood to determine the MLE for μ

$$\begin{aligned}
\frac{\partial \ell}{\partial \mu} &= \frac{\partial}{\partial \mu} \sum_{i=1}^{N} -\log(\sqrt{2\pi}) - \log(\sigma) - \frac{(x^{(i)} - \mu)^2}{2\sigma^2} \\
&= \frac{\partial}{\partial \mu} \sum_{i=1}^{N} -\frac{(x^{(i)} - \mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{N} \frac{1}{2\sigma^2} 2(x^{i)} - \mu) \\
\sum_{i=1}^{N} \frac{1}{2\sigma^2} 2(x_i - \mu) &= 0 \\
\sum_{i=1}^{N} x_i - \mu &= 0 \\
\sum_{i=1}^{N} \mu &= \sum_{i=1}^{N} x_i \\
N\mu &= \sum_{i=1}^{N} x_i \\
\mu &= \frac{\sum_{i=1}^{N} x_i}{N}
\end{aligned}$$

Hence $\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}$, or the average as expected

4.3 Multivariate Gaussian

Assume you observe $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, where each $\mathbf{x}^{(i)} \in \mathbb{R}^M$ is drawn from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Derive the log-likelihood function first.

$\mathbf{x}^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{\mu}, \Sigma$ are unknown. We want to estimate them by maximizing $\mathcal{L}$. p is the pdf function. Note that $\mathcal{L}$ is the likelihood function while ℓ is the log-likelihood function.

$$\mathcal{L}(\boldsymbol{\mu}, \Sigma; \mathbf{x}) = \prod_{i=1}^{N} p(\mathbf{x}^{(i)} | \boldsymbol{\mu}, \Sigma)$$

$$\ell(\boldsymbol{\mu}, \Sigma; \mathbf{x}) = \log \prod_{i=1}^{N} \frac{1}{(2\pi)^{M/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}) \right)$$

$$= \sum_{i=1}^{N} \left(-\frac{M}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}) \right)$$

$$\ell(\boldsymbol{\mu}, \Sigma; \mathbf{x}) = -\frac{NM}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu})$$

Now, using this result, derive $\hat{\boldsymbol{\mu}}$.

note that $\frac{\partial}{\partial \mathbf{v}} \mathbf{v}^T A \mathbf{v} = 2A\mathbf{v}$ if A is symmetric

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{N} \Sigma^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}) = 0$$

Since Σ is positive definite, $0 = N\boldsymbol{\mu} - \sum_{i=1}^{N} \mathbf{x}^{(i)}$

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} = \bar{\mathbf{x}} \text{ as expected}$$

5 Colab Demo Link