1 Probability Definitions

Data in bioinformatics is noisy. It can be due to measurement noise or errors in computation. For instance, for efficiency we sometimes compute in a frequency transformed space after computing a Fourier transform of a timed signal. This would leads to a propagation of errors of the input timed signal samples to the transformed frequency samples. In another scenario, supposed we want to accurately estimate the effective dosage rate of a drug, given estimates of the drug's maximum binding affinity to the disease molecule. This translates to solving a stochastic optimization problem, so that we are be able to tell how close our predicted dosage rate solution is to the true effective rate. In all these scenarios, we regard the input as random variables with certain first and second (and possibly higher) moments (viz, mean, variance, k^{th}-moment) of the input distribution. For example, (and henceforth) we can consider each data pixel of a 2D-image of light intensities (scalar, or vector, ..) acquired through a transmitted or reflected light collection stochastic process, is considered a random variable, having some k moment estimates of its distribution. We can then track the propagation of errors and uncertainty in all the image processing algorithm. Useful stochastic techniques include probabilistic machine learning, expectation maximization etc., many using spectral properties and inequalities for noisy vectors, matrices and tensors.

1.1 Random Variable

A *random variable* (r.v.) X is values as result of an outcome. A *sample space* is a set of all possible outcomes. $\Pr[x] \in [0,1]$ is the probability of occurrence of each outcome. A function that assigns probabilities is called a *probability distribution function*. For example, the uniform distributions, the Gaussian distributions (which have nice concentration properties), and the Poisson distributions (which often appear in image pixels because they count the number of hits over time).

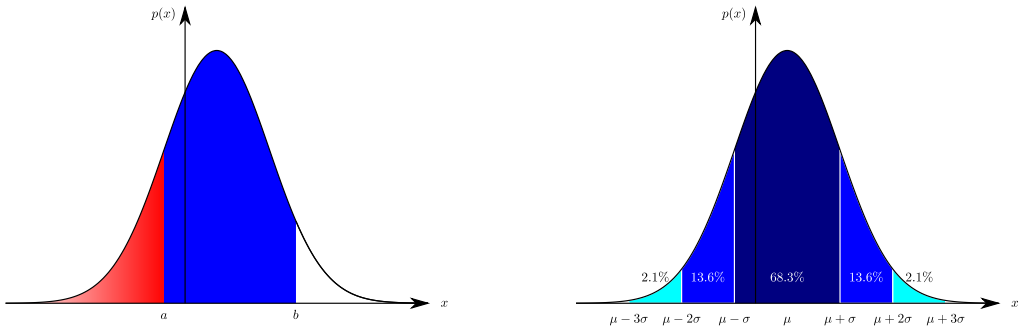
Sometimes, probability mass function does not exists. For example, if X has uniform probability in $[0,1]$, then $\Pr[X = 0.527] = \frac{1}{\infty} = 0$. We can define *probability density function* (pdf) p such that

$$\Pr[a \leq X \leq b] = \int_a^b p(x)\ \mathrm{d}\,x.$$

Notice that the integral is a linear operator on p. The *cummulative distribution function* (cdf) is

$$P(a) = \Pr[X \leq a] = \int_{-\infty}^a p(x)\ \mathrm{d}\,x.$$

Figure 1(a) shows the integrals as areas under the probability density function.



(a) (b)

Figure 1: (a) The probability $\Pr[a \leq X \leq b] = \int_a^b p(x)\ \mathrm{d}\,x$ and the cummulative probability $P(a) = \Pr[X \leq a] = \int_{-\infty}^a p(x)\ \mathrm{d}\,x$ as areas under the probability density function $p(x)$. (b) Approximately 68.3%, 95.4% and 99.7% of probability mass of a Gaussian within σ, 2σ and 3σ from the mean μ.

An *event* is a subset of the sample space. For example, if we have n unbiased coin flips giving random variables $x_1, x_2, \ldots, x_n \in \{0,1\}$, then the sample space is $\{0,1\}^n$. The event of an odd number of ones occurring in the sequence consists of elements in $\{0,1\}^n$.

1.2 Independence

For two events A and B, we define the *conditional probability* as

$$\Pr[A|B] = \Pr[A \cap B] / \Pr[B],$$

where $\Pr[A \cap B]$ is the probability of the joint occurrence of the events.

We say that two events A and B are *independent* if $\Pr[A \cap B] = \Pr[A] \Pr[B]$, or equivalently $\Pr[A|B] = \Pr[A]$. A sequence of n random variables $x_1, x_2, \ldots, x_n$ are *mutually independent* if for all possible $A_1, A_2, \ldots, A_n$, of values of $x_1, x_2, \ldots, x_n$,

$$\Pr[x_1 \in A_1, x_2 \in A_2, \ldots, x_n \in A_n] = \Pr[x_1 \in A_1] \Pr[x_2 \in A_2] \ldots \Pr[x_n \in A_n].$$

Notice that pairwise independence (or even k-wise independence) is weaker than mutual independence.

1.2.1 Bayes Theory

We will learn to use probability theory and sampling for parameter estimation. Consider the Bayes rule.

$$\Pr[A|B] = \frac{\Pr[A] \Pr[B|A]}{\Pr[B]}$$

This follows from $\Pr[A|B] \Pr[B] = \Pr[B|A] \Pr[A]$. We can regard B as the measurement samples that we have. Using this data, we try to estimate A. In the numerator, $\Pr[B|A]$ is the likelihood of A and $\Pr[A]$ is the prior probability. The normalization appears in the denominator $\Pr[B]$. The left hand side is the posterior probability $\Pr[A|B]$.

For example, suppose that a product is defective 0.1% of the time, and a test fails 1% of the time to detect a defective product. Also, assuming that a product is not defective, a test says a product is defective 2% of the time.

Let A be the event that a product is defective. Let B be the event that a test says a product is defective. Then, we have the followings.

$$\begin{aligned}
\Pr[B|A] &= 0.99 \\
\Pr[A] &= 0.001 \\
\Pr[B|\overline{A}] &= 0.02 \\
\Pr[B] &= \Pr[B|A] \Pr[A] + \Pr[B|\overline{A}] \Pr[\overline{A}] = 0.99 \times 0.001 + 0.02 \times 0.999 = 0.02097
\end{aligned}$$

So, using Bayes rule, we can estimate

$$\Pr[A|B] = \frac{\Pr[B|A] \Pr[A]}{\Pr[B]} = \frac{0.99 \times 0.001}{0.0297} \approx 0.047,$$

This estimate is of course, very surprising.

Bayes rule can also be applied to molecule reconstruction from projection images (along with many other example scenarios). In such applications, the multiple molecule image samples are used to reconstruct the structure of the molecule. This is analogous to using measurements to make estimations according to Bayes rule.

1.3 Expectation

The *expectation* of a random variable X with pdf p is defined as

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x p(x) \, \mathrm{d}\, x.$$

The linearity of expectation

$$\mathrm{E}[X_1 + X_2 + \ldots + X_n] = \mathrm{E}[X_1] + \mathrm{E}[X_2] + \ldots + \mathrm{E}[X_n]$$

holds even without independence.

The union bound

$$\Pr[A_1 \cup A_2 \cup \ldots \cup A_n] \leq \sum_{i=1}^{n} \Pr[A_i]$$

is an upper bound of the unions of events.

The inclusion-exclusion principle says that

$$\Pr[A_1 \cup A_2 \cup \ldots \cup A_n] = \sum_{i=1}^{n} \Pr[A_i] - \sum_{i<j} \Pr[A_i \cap A_j] + \sum_{i<j<k} \Pr[A_i \cap A_j \cap A_k] - \ldots.$$

One application of the inclusion-exclusion principle is volume calculation of molecules represented as a union of atoms. An atom consists of a nucleus in its center surrounded by a electron cloud, which can be represented by a ball with radius equal to the range of its van der Waals force. The atoms are bonded together, forming a geometry of union of balls. Examples include NaCl salt, protein, and water molecule (H_2O) which polarizes like a magnet with Hydrogen (H) positively charged and Oxygen (O) negatively charged. Two (or a small number of) balls may overlap each other. Since the volume is proportional to finding electrons in certain region, we can apply the inclusion-exclusion principle.

1.4 Variance

The *variance* of a random variable $X \in \mathbb{R}$ is given by

$$\begin{aligned} \mathrm{Var}(X) = \sigma^2(X) &= \mathrm{E}[X - \mathrm{E}^2[X]]^2 \\ &= \mathrm{E}[X^2] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + \mathrm{E}^2[X] \\ &= \mathrm{E}[X^2] - \mathrm{E}^2[X]. \end{aligned}$$

For a Gaussian random variable, its standard deviation σ tells that more than 68% of the probability mass is with σ from its mean. For 2σ and 3σ from the mean, the probability masses are more than 95% and 99% respectively. (Figure 1(b))

In general, $\mathrm{Var}(X_1 + X_2) \neq \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$. However, equality holds if X_1 and X_2 are pairwise independent. In fact, if $X_1, X_2, \ldots, X_n$ are pairwise independent, then

$$\mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

1.5 Markov and Chebyshev inequalities and Law of Large Numbers

The following probability inequalities by Markov and Chebyshev are used to prove the **Law of Large Numbers**.

Theorem (Markov Inequality [BHK, Theorem 2.1] [MU, Theorem 3.1])**.** *Let $X \geq 0$ be a random variable and $a > 0$.*

$$\Pr[X \geq a] \leq \mathrm{E}[X]/a.$$

Proof. For a continuous non-negative random variable x with probability density $p(x)$

$$\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty xp(x)\ \mathrm{d}\,x \\
&= \int_0^a xp(x)\ \mathrm{d}\,x + \int_a^\infty xp(x)\ \mathrm{d}\,x \\
&\geq \int_a^\infty xp(x)\ \mathrm{d}\,x \\
&\geq \int_a^\infty ap(x)\ \mathrm{d}\,x \\
&= a\int_a^\infty p(x)\ \mathrm{d}\,x \\
&= a\,\mathrm{Pr}[X \geq a]
\end{aligned}$$

□

Note the proof works for discrete probability distributions. Replace summations for all the integrals.

Corollary. *Let X be a non-negative random variable and $c > 0$. Then,* $\mathrm{Pr}[X \geq c\,\mathrm{E}[X]] \leq 1/c$.

The above says that the value of X is not far from the mean $E(X)$. The Chebyshev's inequality (below) can be proved by applying Markov's inequality on the variance.

Theorem (Chebyshev Inequality [BHK, Theorem 2.3] [MU, Corollary 3.7])**.** *Let X be a random variable with mean m and variance σ^2. Then, for all $a > 0$,*

$$\mathrm{Pr}[\,|X - m| \geq a\sigma\,] \leq 1/a^2.$$

Proof.

$$Pr(|X - m| \geq c) \Rightarrow Pr(|X - m|^2 \geq c^2)$$

By setting $c = a\sigma$ and apply Markov Inequality, one have:

$$Pr(|X - m|^2 \geq a^2\sigma^2) \leq \frac{E(|X - m|^2)}{a^2\sigma^2} = \frac{\sigma^2}{a^2\sigma^2} = \frac{1}{a^2}$$

□

Using the Chebyshev inequality, we can now prove the Law of Large Numbers.

Theorem (**Law of Large Numbers).** *Let S be the sample mean of n independent random variables $X_1, X_2, \ldots, X_n$ with means* $\mathrm{E}[X_i] = m$ *and variances* $\mathrm{Var}[X_i] = \sigma^2$. *Then, for all $\varepsilon > 0$,*

$$\mathrm{Pr}[|S - m| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Proof. Applying Chebyshev Inequality on S with $a = \varepsilon/\sigma(S)$, we get

$$\begin{aligned}
\mathrm{Pr}\,[|S - m| \geq \varepsilon] &\leq \sigma^2(S)/\varepsilon^2 \\
&= \frac{1}{\varepsilon^2}\sigma^2\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) \\
&= \frac{1}{n^2\varepsilon^2}\sigma^2(X_1 + X_2 + \ldots + X_n) \\
&= \frac{\sigma^2}{n\varepsilon^2}.
\end{aligned}$$

□

2 Probability Distributions

2.1 Gaussian Distributions

The *Gaussian distribution* is related to Central Limit Theorem.

Theorem (Central Limit Theorem [BHK, Theorem 12.2]). *If $X_1, \ldots, X_n \in \mathbb{R}$ is a sequence of independent identically distributed (i.i.d.) random variables each with mean μ and variance σ^2, then*

$$X = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - n\mu\right)$$

converges to the distribution $\mathrm{N}(0, \sigma^2)$.

The univariate Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ is given by the pdf

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

For d-variate Gaussian distribution $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean vector and $\Sigma \in \mathbb{R}^{d\times d}$ is the *covariance matrix*, the pdf is given by

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

Recall the unit vector $\mathbf{v}$, is with $\|\mathbf{v}\| = 1$. Now give a point $\mathbf{x}$ and the corresponding unit vector $\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$ and $\cos\phi = \langle\hat{\mathbf{x}}, \mathbf{v}\rangle$. The position of the projected point $P_{\mathbf{x}} = \|\mathbf{x}\| \cdot \cos\phi = \|\mathbf{x}\| \cdot \langle\hat{\mathbf{x}}, \mathbf{v}\rangle = \langle\mathbf{x}, \mathbf{v}\rangle$ and the projected vector is given by $\langle\mathbf{x}, \mathbf{v}\rangle\mathbf{v}$.

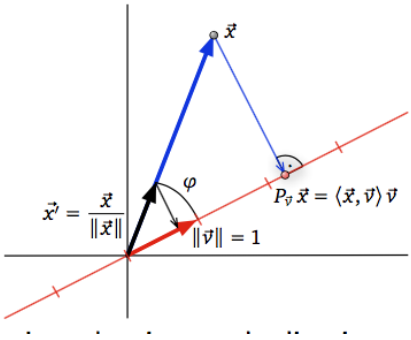
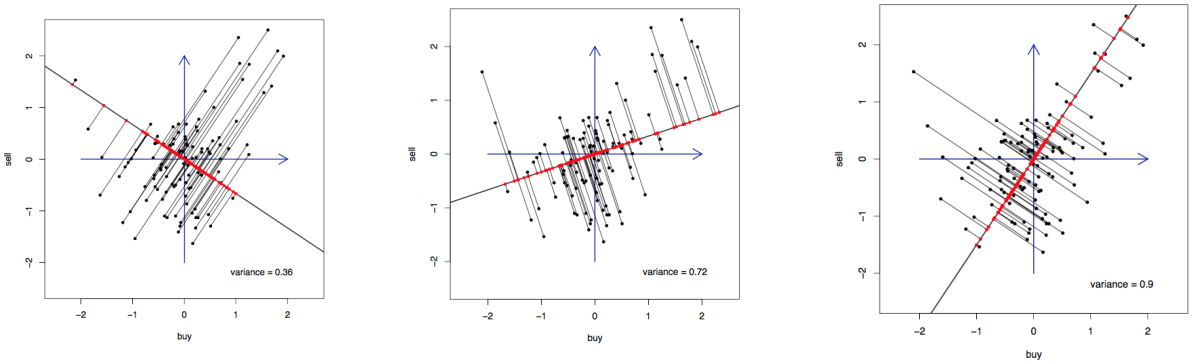


Figure 2: Vectorizing and visualizing Projections.

The $1D$ variance of the data projected onto the vector $\mathbf{v}$, is given by $\sigma^2_{\mathbf{v}} = \frac{1}{k-1}\sum_{i=1}^{k}\langle\mathbf{x_i}, \mathbf{v}\rangle^2$. We want to find the direction $\mathbf{v}$ with maximal $\sigma^2_{\mathbf{v}}$. This can be achieved via the calculation of $\sigma^2_{\mathbf{v}}$

$$\begin{aligned}
\sigma^2_{\mathbf{v}} &= \frac{1}{k-1}\sum_{i=1}^{k}\langle\mathbf{x}_i, \mathbf{v}\rangle^2 \\
&= \frac{1}{k-1}\sum_{i=1}^{k}\mathbf{v}^T(\mathbf{x}_i\mathbf{x}_i^T)\mathbf{v} \\
&= \mathbf{v}^T\left(\frac{1}{k-1}\sum_{i=1}^{k}\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{v} \\
&= \mathbf{v}^T C\mathbf{v}_1
\end{aligned} \tag{1}$$



(a) (b) (c)

Figure 3: Linear Regression Fits (a) variance= 0.36 (b) variance= 0.72 (c) variance= 0.9. The best approximation result is given by solving least square minimization problem with respect to the form: $\|Ax - b\|$.

where $C := \left(\frac{1}{k-1}\sum_{i=1}^{k} \mathbf{x}_i \mathbf{x}_i^T\right)$ is called the **covariance matrix** of data points. (C is a $n \times n$ square matrix). The original variance of the data is given by $\sigma^2 = \textbf{trace}(C) = C_{11} + C_{22} + \ldots + C_{nn}$

$$C = \begin{bmatrix} \sigma_1^2 & C_{12} & \cdots & C_{1n} \\ C_{21} & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & C_{n-1,n} \\ C_{n1} & \cdots & C_{n,n-1} & \sigma_n^2 \end{bmatrix}$$

When $d = 3$, an *isotropic* Gaussian has 4 degrees of freedom, corresponding to the number of parameters necessary to define a sphere in $\mathbb{R}^3$. Meanwhile, an *anisotropic* Gaussian would have $9 = \binom{2+3}{2} - 1$ degrees of freedom, corresponding to the number of parameters necessary to define an ellipsoid. If the ellipsoid is *isothetic*, then the degree of freedom reduces to 6.

2.2 Binomial Distributions

A *Bernoulli distribution* is a stochastic process with two outcomes

$$X = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1-p \end{cases}.$$

The *binomial distribution* $\text{Bin}(n, p)$ counts the number X of ones in n independent Bernoulli trials.

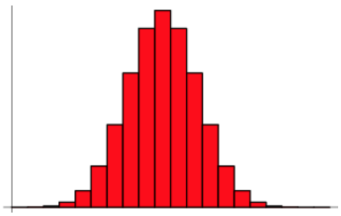
$$\Pr[X = k] = \Pr[(\text{total number of ones}) = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

It has mean np and variance $np(1-p)$. It also satisfies the property that if $X \sim \text{Bin}(n_1, p)$ and $Y \sim \text{Bin}(n_2, p)$ are independent, then $X + Y \sim \text{Bin}(n_1 + n_2, p)$.

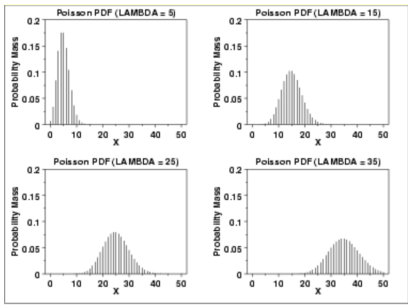
2.3 Poisson Distribution

Let λ be the average rate per unit of time and n be the number of division of a unit time interval into segments, where the probability of two events occurring in the same segment is negligible. The Poisson distribution counts the number X of events occurring in a unit of time as $n \to \infty$. It is the limit of $\text{Bin}(n, p = \lambda/n)$. For the Poisson distribution both the mean and variance equal λ.

$$\Pr[X = k] = \Pr[k \text{ events occurs in a unit of time}] = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$



(a)



(b)

Figure 4: (a) The above plot shows the Binomial distribution of exactly n successes out of N = 20 trials with p = q = $\frac{1}{2}$. (b) The above plot shows the Poisson distribution for four different values of λ.

3 Maximum Likelihood Estimator

Suppose a probability distribution of a random variable X depends on parameter r. So, $\Pr[X|r]$ denotes the probability of observing X if parameter value is r. If r is also random, after observing the value of X, one can find the best r maximizing the posterior probability

$$\Pr[r|X] = \frac{\Pr[X|r]\Pr[r]}{\Pr[X]}.$$

Assume $\Pr[r]$ is the same for all r. Since the unconditional probability of X in the denominator is independent of r, it reduces to finding the *maximum likelihood estimator* (MLE)

$$\arg\max_r \mathrm{L}(r|X) = \arg\max_r \Pr[X|r].$$

Example

Consider the example of flipping a biased coin in n trials with unknown probability r of getting head. The probability of getting k heads follows the binomial distribution $\mathrm{Bin}(n, r)$ such that

$$\Pr[k|r] = \binom{n}{k} r^k (1-r)^{n-k}.$$

If we get 62 heads and 38 tails in 100 trails, the maximum likelihood estimator gives $r = 0.62$ when $\Pr[62|r]$ is maximized. One can see this by setting the derivative (with respect to r) to 0.

We can study a single particle using cryo-electron microscopy. To do so, we build a specimen grid of millions of in-vitro molecules and take a snapshot by shooting X-ray and measuring its projection. We can reconstruct the locations of the molecules by solving a least square optimization with regularizer, but it is unstable. How many samples would we need? We want to show that the solution converges as the sample size increase. We can recast the problem by regarding the data as a random variable with certain mean and variance. We can then solve for the maximum a-posterior estimator. We would also like to output a confidence level of our estimation.

4 Noisy Data Regression

Given data $\mathbf{x}_i \in \mathbb{R}^d$ and observation $y_i \in \mathbb{R}$, suppose we apply the regression model

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

to fit the data, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the Gaussian noise. The way to formulate the optimization problem is to treat the model as a parameter estimation problem: y_i is drawn from a one-dimensional Gaussian distribution with mean $\mathbf{w}^T \mathbf{x}_i$ and variance σ^2, namely:

$$P(y_i|\mathbf{x}_i,\mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mathbf{x}_i^T\mathbf{w}-y_i)^2}{2\sigma^2}}$$

Therefore, we can estimate $\mathbf{w}$. First, consider MLE (Maximum Likelihood Estimator) and log likelihood problem.

$$\begin{aligned}
\mathbf{w} &= \underset{\mathbf{w}}{argmax} \quad \sum_{i=1}^{n} \log(P(y_i \mid \mathbf{x}_i,\mathbf{w})) \\
&= \underset{\mathbf{w}}{argmax} \quad \sum_{i=1}^{n} \left[\log(\frac{1}{\sqrt{2\pi\sigma^2}}) + \log(e^{-\frac{(\mathbf{x}_i^T\mathbf{w}-y_i)^2}{2\sigma^2}})\right] \\
&= \underset{\mathbf{w}}{argmax} \quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w}-y_i)^2 \\
&= \underset{\mathbf{w}}{argmin} \quad \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w}-y_i)^2
\end{aligned} \tag{2}$$

Let us denote $L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w}-y_i)^2$ as the loss function. Then it will automatically become the original least square problem as people generally did. And we know the solution of the least square problem is:

$$\mathbf{w} = X^{\dagger}\mathbf{y}^T \quad \text{where } X = [\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n], \mathbf{y} = [y_1,y_2,\ldots,y_n]$$

Here $X^{\dagger}$ represents the pseudo-inverse of matrix. Next, we consider maximum a posteriori probability (MAP) estimate. This will apply Bayes' theorem. Assume we have prior model:

$$P(\mathbf{w}) = \frac{1}{\sqrt{2\pi\alpha^2}} e^{-\frac{\mathbf{w}^T\mathbf{w}}{2\alpha^2}}$$

which implies that $\mathbf{w}_i \sim \mathcal{N}(0,\alpha)$, then we estimate w by calculating posterior distribution of $\mathbf{w}$:

$$\begin{aligned}
\mathbf{w} &= \underset{\mathbf{w}}{argmax} \quad P(\mathbf{w} \mid y_1,\mathbf{x}_1,y_2,\mathbf{x}_2,\ldots,y_n,\mathbf{x}_n) \\
&= \underset{\mathbf{w}}{argmax} \quad \frac{P(y_1,y_2,\ldots,y_n \mid \mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\mathbf{w})P(\mathbf{w})}{P(y_1,y_2,\ldots,y_n)} \\
&= \underset{\mathbf{w}}{argmax} \quad P(y_1,y_2,\ldots,y_n \mid \mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\mathbf{w})P(\mathbf{w}) \\
&\text{(Since } P(y_1,y_2,\ldots,y_n) \text{ is a constant.)} \\
&= \underset{\mathbf{w}}{argmax} \quad P(y_1 \mid y_2,\ldots,y_n,\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\mathbf{w})P(y_2 \mid y_3,\ldots,y_n,\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\mathbf{w}) \\
&\qquad \cdots P(y_n \mid \mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\mathbf{w})P(\mathbf{w}) \\
&\text{(By independence)} \\
&= \underset{\mathbf{w}}{argmax} \quad \left[\prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i,\mathbf{w})\right] P(\mathbf{w})
\end{aligned} \tag{3}$$

Then consider the logarithm of the function:

$$\begin{aligned}
\mathbf{w} &= \underset{\mathbf{w}}{argmax} \quad \sum_{i=1}^{n} \ln P(y_i \mid \mathbf{x}_i,\mathbf{w}) + \ln P(\mathbf{w}) \\
&= \underset{\mathbf{w}}{argmax} \quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w}-y_i)^2 - \frac{1}{2\alpha^2}\mathbf{w}^T\mathbf{w}_i \\
&= \underset{\mathbf{w}}{argmin} \quad -\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w}-y_i)^2 + \lambda\|\mathbf{w}\|_2^2
\end{aligned} \tag{4}$$

This is known as Ridge Regression problem, and the closed form of the solution is :

$$\mathbf{w} = (XX^T + \lambda I)^{-1} X\mathbf{y}^T \quad \text{where } X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n], \mathbf{y} = [y_1, y_2, \ldots, y_n]$$

5 Unbiased Estimator

Let $X = (X_1, X_2, \ldots, X_n)$ be samples or observations from a distribution having parameter θ. (For example, the Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ has parameters mean μ and variance σ^2, while the binomial distribution $\mathrm{Bin}(n, p)$ has parameters n and success probability p.)

Let $D(X)$ be an estimator of some function $h(\theta)$. The *bias* is defined as

$$\mathrm{E}[D(X) - h(\theta)].$$

It is called an *unbiased estimator* when the bias equals zero.

The quality of the estimator can be measured by the *mean squared error* (MSE)

$$\mathrm{E}[(D(X) - h(\theta))^2] = \mathrm{Var}(D(X)) + \mathrm{Bias}^2.$$

Theorem. *Let $X_1, X_2, \ldots, X_n$ be independent samples, each with mean μ and variance σ^2.*

1. *[?, Example 14.3] $D(X) = \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator of μ.*
2. *If μ is known, then $D(X) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ is an unbiased estimator of σ^2.*
3. *[?, Example 14.5] If μ is not known, then $D(X) = \frac{1}{n-1}\sum_{i=1}^n (X_i - m)^2$ is an unbiased estimator of σ^2, where $m = \frac{1}{n}\sum_{i=1}^n X_i$.*

Proof of 3. Let $S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - m)^2$. Observe that

$$\begin{aligned}\sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n \Big[(X_i - m) + (m - \mu)\Big]^2 \\ &= \sum_{i=1}^n (X_i - m)^2 + n(m - \mu)^2\end{aligned}$$

Hence,

$$\begin{aligned}S^2 &= \frac{1}{n}\sum_{i=1}^n (X_i - m)^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 - (m - \mu)^2 \\ \mathrm{E}[S^2] &= \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(X_i) - \mathrm{Var}(m) \\ &= \frac{n-1}{n}\sigma^2\end{aligned}$$

Thus, $\left(\frac{n}{n-1}S^2\right)$ is an unbiased estimator of σ^2. □

References

[BHK] Avrim Blum, John Hopcroft and Ravindran Kannan. *Foundations of Data Science.*

[BM] George Edward Pelham Box and Mervin Edgar Muller. *A Note on the Generation of Random Normal Deviates.* The Annals of Mathematical Statistics, 1958. Vol. 29, No. 2, pp. 610–611.

[MU] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press, 2005.