ECE 469/ECE 568 - Machine Learning

Southern Illinois University

September 23, 2024

Example: Gaussian Random variable

- The Gaussian random variables are commonly encountered in statistical analysis of machine learning.
- The PDF of a Gaussian variable is completely characterized by its mean μ_X and variance σ_X^2.
- Then, the PDF of a Gaussian random variable is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right)$$

- In machine learning, the prior PDF of unknown variables are modeled by using Gaussian distribution.

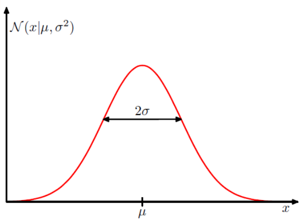
Example: Gaussian Random variable

- We denote that the random variable X is Gaussian distributed with mean μ_X and variance σ_X^2 as

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

- The square-root of the variance is termed as the standard deviation.
- If $X \sim \mathcal{N}(0, 1)$, then X is a standard Gaussian random variable.
- A Gaussian random variable $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ can be normalized to obtain a standard Gaussian random variable Z as $Z = (X - \mu_X)/\sigma_X^2$.

Example: Gaussian Random variable



- The area under a PDF $f(x)$ is always one.

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\text{mean} = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \mu$$

$$\text{variance} = \mathbb{E}[(X-\mu)^2] = \sigma^2$$

Joint moments

- Let X and Y be two random variables. The correlation is defined as

$$\text{corr}[XY] = \mathbb{E}[XY]$$

- The correlation of the centered random variables $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ is called the covariance of X and Y.

$$cov[XY] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

- By letting $\mathbb{E}[X] = \mu_x$ and $\mathbb{E}[Y] = \mu_Y$, $cov[XY]$ can be expressed as

$$\begin{aligned} cov[XY] &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= \mathbb{E}[XY] - \mu_x \mu_Y \end{aligned}$$

- Let σ_X^2 and σ_Y^2 denote the variances of X and Y, respectively. The covariance of X and Y, normalized with respect to $\sigma_X \sigma_Y$ is called the correlation coefficient:

$$\rho = \frac{cov(XY)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{cov(XY)}{\sigma_X \sigma_Y}$$

Independent, Uncorrelated and Orthogonal random variables

- Two random variables X and Y are statistically independent if and only if the joint probability density function equals to the product of their marginal densities:

$$\boxed{f_{X,Y}(x,y) = f_X(x)f_Y(y) \qquad \Longrightarrow \qquad \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]}$$

- Two random variables X and Y are uncorrelated if and only if their covariance ($cov(XY) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$) is zero; that is if and only if

$$cov[XY] = 0$$

- Two random variables X and Y are orthogonal if their correlation is zero; that is if and only if

$$\mathbb{E}[XY] = 0$$

- If X and Y are statistically independent, then they are uncorrelated. However, the converse of this statement is not necessarily true.
- Nevertheless, for Gaussian variables, uncorrelated-ness indeed implies statistically independence.

Central Limit Theorem

- The central limit theorem provides the mathematical justification for using a Gaussian process as a model for a large number of different physical phenomena in which the observed random variable is a result of a large number of individual random events.
- To formulate this important theorem, let X_i for $i \in \{1, 2, \cdots, N\}$ be a set of random variables that satisfies the following requirements:
  1. The X_i are statistically independent
  2. The X_i have the same probability distribution with mean μ_X and variance σ_X^2

Central Limit Theorem Continued...

- The $X_i's$ so described are said to continue a set of independently and identically distributed (i.i.d.) random variables. Let these random variables be normalized as follows:

$$Y_i = \frac{X_i - \mu_X}{\sigma_X}, \qquad \text{for} \qquad i = 1, 2, \cdots, N$$

Therefore, Y_i's have zero mean and unit variance.

- Let us define a new random variable V_N as follows:

$$V_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Y_i$$

- The central limit states that the probability distribution of V_N approaches a normalized Gaussian distribution with zero mean and unit variance in the limit as N approaches infinity.

$$\lim_{N \to \infty} V_N \sim \mathcal{N}(0, 1)$$

The multivariate Gaussian distribution

- A multivariate Gaussian distribution is completely characterized by its mean vector ($\boldsymbol{\mu}$) and the covariance matrix ($\mathbf{C}$).
- The Gaussian PDF of a N-dimensional vector ($\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$) is

$$f_X(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\mathbf{C}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

- Here $\boldsymbol{\mu}$ is the mean vector $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$.
- Moreover, $\mathbf{C}$ is the covariance matrix defined by $\mathbf{C} = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$.
- $|C| = \det(\mathbf{C})$ is the determinant of the covariance matrix.

The likelihood function

- In parameter estimation, the PDF is typically called the likelihood function.
- For example, assume that a data set of observations $\mathbf{x} = (x_1, \cdots, x_N)^T$ is drawn independently and identically (i.i.d.) from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.
- The likelihood function of the data set $\mathbf{x}$ is

$$f_{\mathbf{x}}(\mathbf{x}) = \prod_{n=1}^{N} f_{X_n}(x_n) = \prod_{n=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_n - \mu)^2}{2\sigma^2} \right) \right)$$

- The likelihood function can further be expanded as

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right)$$

- The log-likelihood function is computed by taking the $\log_e(\cdot)$ in both sides of likelihood function:

$$\ln\left(f_{\mathbf{x}}(\mathbf{x})\right) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln(\sigma^2) - \frac{N}{2} \ln(2\pi)$$

Maximum likelihood estimation

- Unknown parameters can be estimated by maximizing the underlying likelihood function –> maximum likelihood estimation (MLE)
- For example, assume that a data set of observations $\mathbf{x} = (x_1, \cdots, x_N)^T$ has i.i.d. Gaussian $\mathcal{N}(\mu, \sigma^2)$ entries with an unknown mean μ. We can estimate μ using the maximum likelihood estimation technique.
- Maximum likelihood technique involves maximizing the PDF over the unknown parameter μ.

$$\hat{\mu} = \mathrm{argmax}_{\mu} f_{\mathbf{x}}(\mathbf{x})$$

- Since $\log_e(\cdot)$ or $\ln(\cdot)$ is an increasing function of its argument, we can maximize the log-likelihood function.

$$\hat{\mu} = \mathrm{argmax}_{\mu} \ln\left(f_{\mathbf{x}}(\mathbf{x})\right)$$

- By substituting $\ln\left(f_{\mathbf{x}}(\mathbf{x})\right)$ we have

$$\hat{\mu} = \mathrm{argmax}_{\mu} \left(\ln\left(f_{\mathbf{x}}(\mathbf{x})\right) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln(\sigma^2) - \frac{N}{2} \ln(2\pi) \right)$$

Maximum likelihood estimation

- We can omit the last two terms on the right-hand side because they do not depend on μ.
- Also note that scaling the log likelihood by a positive constant coefficient does not alter the location of the maximum with respect to $\mathbf{w}$.
- Instead of maximizing the log likelihood, we can equivalently minimize the negative log likelihood.

$$\hat{\mu} = \operatorname{argmin}_{\mu} \left(\sum_{n=1}^{N} (x_n - \mu)^2 \right)$$

- Thus, we can see that maximizing likelihood is equivalent (determining $\mathbf{w}$ is concerned) to minimizing the sum-of-squares error function.
- The sum-of-squares error function has resulted as a consequence of maximizing likelihood under the assumption of a Gaussian noise distribution.

Maximum likelihood estimation

- The stationary point of this minimization can be found as

$$\begin{aligned}\frac{d}{d\mu}\ln\left(f_{\mathbf{x}}(\mathbf{x})\right) &= 0\\ \sum_{n=1}^{N}(x_n-\mu) &= 0\\ \hat{\mu} &= \frac{1}{N}\sum_{n=1}^{N}x_n\end{aligned}$$

- Thus, the maximum likelihood of the mean is actually the the sample mean or the mean of the observed values $\{x_n\}$.
- Similarly, by maximizing $\ln\left(f_{\mathbf{x}}(\mathbf{x})\right)$ with respect to σ^2, we obtain the maximum likelihood estimation for the variance as

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

- Note that this is also the sample variance measured with respect to the sample mean $\hat{\mu}$.

Maximum likelihood estimation

- The mean of the maximum likelihood of the mean ($\hat{\mu}$) is given by

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] = \frac{1}{N}\sum_{n=1}^{N}\mu = \mu$$

- On average the maximum likelihood estimate obtains the true mean.
- However, the mean of the estimate of the variance ($\hat{\sigma^2}$) is given by

$$\mathbb{E}[\hat{\sigma^2}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2\right] = \left(\frac{N-1}{N}\right)\sigma^2$$

- Hence, on average the maximum likelihood estimate underestimates the true variance by a factor $(N-1)/N$.

Maximum likelihood estimation

- The maximum likelihood estimation underestimates the variance of the distribution. This phenomenon is called the bias effect, and related to the problem of over-fitting in polynomial curve fitting.
- However, we can modify the variance estimator to be unbiased as follows:
$$\tilde{\sigma^2} = \frac{N}{N-1}\hat{\sigma^2} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$
- However, the bias of the maximum likelihood estimator of the variance becomes insignificant when $N \to \infty$.

Bias of an estimator

- Bias of an estimator measures the expected/average deviation from the true value of the function or parameter.
- The bias of an estimator ($\hat{x}$) is defined as

$$\mathrm{bias}(\hat{x}) = \underbrace{\mathbb{E}[\hat{x}]}_{\text{mean of the estimator}} - \underbrace{x}_{\text{true value}}$$

- An estimator $\hat{x}$ is said to be unbiased if $\mathrm{bias}(\hat{x}) = 0$

$$\text{unbiased estimator:} \quad \mathbb{E}[\hat{x}] = x.$$

- An estimator is asymptotically unbiased if

$$\lim_{N \to \infty} \mathrm{bias}(\hat{x}) = 0$$

where N is the number of samples or sample size.

Variance of an estimator

- Variance of an estimator provides a measure of the deviation from the expected estimator value that any particular sampling of the data is likely to cause.
- The variance of the estimator $\hat{x}$ is

$$\mathrm{Var}[\hat{x}] = \mathbb{E}\left[(\hat{x} - \mathbb{E}[\hat{x}])^2\right]$$

- In machine learning, the square-root of the variance of an estimator is called the standard error.

$$\mathrm{SE}(\hat{x}) = \sqrt{\mathrm{Var}[\hat{x}]}$$

- We prefer estimators with low variances or low standard errors.

Mean squared error (MSE) of an estimator

- The mean squared error (MSE) of an estimator is given by

$$\mathrm{MSE} = \mathbb{E}[(\hat{x} - x)^2]$$

- The MSE can be expressed in terms of the mean and variance of the estimator as

$$\begin{aligned}
\mathrm{MSE} &= \mathbb{E}[(\hat{x} - x)^2] \\
&= \mathbb{E}[((\hat{x} - \mathbb{E}[\hat{x}]) + (\mathbb{E}[\hat{x}] - x))^2] \\
&= \underbrace{\mathbb{E}[(\hat{x} - \mathbb{E}[\hat{x}])^2]}_{\mathrm{Var}[\hat{x}]} + \underbrace{(\mathbb{E}[\hat{x}] - x)^2}_{\mathrm{bias}^2(\hat{x})}
\end{aligned}$$

- Thus we have $\mathrm{MSE} = \mathrm{Var}[\hat{x}] + \mathrm{bias}^2(\hat{x})$.
- The MSE of an unbiased estimator is just the variance of the estimator.
- We would like to find unbiased estimators with minimum variances –> This leads to the concept of "minimum variance unbiased estimators".