Anuran Makur
01/17/2023

Leave-one-out Approximation of Integrated Squared Error:

① Approximating means:

Consider independent and identically distributed r.v.s $X_1, X_2, \dots, X_n$ such that $X_i \in \{-1, 1\}$ and $\mathbb{P}(X_i = 1) = p \in (0,1)$. Then, we define the expected value of X_i as

$$\mathbb{E}[X_i] \triangleq 1 . \mathbb{P}(X_i = 1) + (-1)\, \mathbb{P}(X_i = -1) = p - (1-p) = 2p - 1 .$$

When n is large, $\underbrace{p \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = 1\}}_{\text{proportion of } X_i\text{'s equal to } 1}$ and $1 - p \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = -1\}$.

Hence, when n is large,

$$\begin{aligned}\mathbb{E}[X_i] &\approx 1 . \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = 1\}\right) + (-1) . \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = -1\}\right) \\ &= \frac{1}{n}\left(1 . \{\text{no. of } X_i\text{'s} = 1\} + (-1) . \{\text{no. of } X_i\text{'s} = -1\}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i .\end{aligned}$$

This idea can be generalized. Suppose $X_1, \dots, X_n \in \mathbb{R}$ with probability density function $f(x)$. Then,

$$\mathbb{E}[X_i] = \int_{-\infty}^{+\infty} x\, f(x)\, dx \approx \frac{1}{n}\sum_{i=1}^{n} X_i ,$$

$$\mathbb{E}[g(X_i)] = \int_{-\infty}^{+\infty} g(x) f(x)\, dx \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i) ,$$

when n is large, for any function $g : \mathbb{R} \to \mathbb{R}$.

② Decomposing Integrated Squared Error:

$$L(w) \triangleq \int_{-\infty}^{\infty} \left(\hat{f}_m(x) - f(x)\right)^2 dx = \underbrace{\int_{-\infty}^{\infty} \hat{f}_m(x)^2\, dx - 2\int_{-\infty}^{\infty} \hat{f}_m(x) \overset{\text{unknown!}}{f(x)}\, dx}_{\text{approximate this with } J(w)} + \underbrace{\int_{-\infty}^{\infty} f(x)^2\, dx}_{+ \text{ constant}}$$

③ Leave-one-out Cross-Validation [Rudemo '82]:

We would usually approximate $\int_{-\infty}^{\infty} \hat{f}_m(x) f(x)\, dx$ as

$$\int_{-\infty}^{\infty} \hat{f}_m(x) f(x)\, dx \approx \frac{1}{m}\sum_{i=1}^{m} \hat{f}_m(X_i)$$

where $X_1, \dots, X_m$ are samples from $f(x)$. However, $\hat{f}_m$ depends on X_i, which leads to greater "bias".

So, we use the alternative approximation

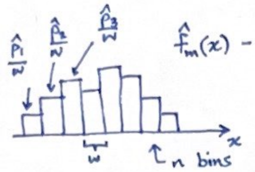
$$\int_{-\infty}^{\infty} \hat{f}_m(x) f(x)\, dx \approx \frac{1}{m}\sum_{i=1}^{m} \hat{f}_{m,\neg i}(X_i) \quad \leftarrow \text{lower "bias"}$$

where $\hat{f}_{m,\neg i}$ is the histogram made of samples $X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_m$ (leave-one-out!), and $\hat{f}_{m,\neg i} \approx \hat{f}_m$ for large m.

Anuran Makur
01/17/2023

④ Approximation $J(w)$:

Recall that:



$\hat{f}_m(x)$ - frequency density

$$\int_{-\infty}^{\infty} \hat{f}_m(x)^2\,dx - 2\int_{-\infty}^{\infty} \hat{f}_m(x) f(x)\,dx \approx \sum_{k=1}^{n} w\left(\frac{\hat{p}_k}{w}\right)^2 - 2\cdot\frac{1}{m}\sum_{i=1}^{m} \overbrace{\hat{f}_{m,-i}(X_i)}^{=\frac{m\hat{p}_k - 1}{w(m-1)}\text{, where } X_i \text{ is in bin } k}$$

$$= \frac{1}{w}\sum_{k=1}^{n}\hat{p}_k^2 - 2\frac{1}{m}\sum_{k=1}^{n} \underbrace{m\hat{p}_k}_{\substack{\text{no. of}\\ X_i\text{'s in bin } k}} \cdot \underbrace{\frac{m\hat{p}_k - 1}{w(m-1)}}_{\text{value of histogram}}$$

$$= \frac{1}{w}\sum_{k=1}^{n}\hat{p}_k^2 - \frac{2m}{w(m-1)}\sum_{k=1}^{n}\hat{p}_k^2 + \frac{2}{w(m-1)}\underbrace{\sum_{k=1}^{n}\hat{p}_k}_{=1}$$

$$= \frac{2}{w(m-1)} - \frac{m+1}{w(m-1)}\sum_{k=1}^{n}\hat{p}_k^2$$

Hence, we define $\boxed{J(w) \triangleq \frac{2}{w(m-1)} - \frac{m+1}{w(m-1)}\sum_{k=1}^{n}\hat{p}_k^2}$ and we have:

$$\boxed{L(w) \approx J(w) + \underbrace{\int_{-\infty}^{\infty} f(x)^2\,dx}_{\text{constant}}}.$$

1) Properties of mean and variance:

i) For random variables X and Y, and constants $a,b,c \in \mathbb{R}$, we have
$$\mathbb{E}[aX+bY+c] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y] + c.$$

ii) For independent random variables X and Y, and constants $a,b,c \in \mathbb{R}$, we have
$$\text{var}(aX+bY+c) = a^2\,\text{var}(X) + b^2\,\text{var}(Y).$$

2) Unbiased estimators of mean and variance:

Let $X_1,\dots,X_n$ be i.i.d. random variables with expected value $\mathbb{E}[X]=\mu$ and variance $\text{var}(X)=\sigma^2$.
Let $\frac{1}{n}\sum_{i=1}^{n} X_i \triangleq \bar{X}$ be the empirical estimate of μ. Then,
$$\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu. \leftarrow \text{unbiased!}$$

Let $s^2 \triangleq \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ be the empirical estimate of σ^2. Observe that:
$$\begin{aligned}
\mathbb{E}\Big[\sum_{i=1}^{n}(X_i-\bar{X})^2\Big] &= \mathbb{E}\Big[\sum_{i=1}^{n} X_i^2 - 2X_i\bar{X} + \bar{X}^2\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\bar{X}^2\Big] \\
&= \sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2n\,\mathbb{E}[\bar{X}^2] + \sum_{i=1}^{n}\mathbb{E}[\bar{X}^2] \\
&= n\,\mathbb{E}[X^2] - n\,\mathbb{E}[\bar{X}^2] \\
&= n\big(\text{var}(X) + \mathbb{E}[X]^2\big) - n\Big(\text{var}\Big(\frac{1}{n}\sum_{i=1}^{n}X_i\Big) + \mathbb{E}[\bar{X}]^2\Big) \\
&= n\sigma^2 + \cancel{n\mu^2} - n\Big(\frac{\text{var}(X)}{n} + \cancel{\mu^2}\Big) \\
&= (n-1)\sigma^2.
\end{aligned}$$

Hence, $\mathbb{E}[s^2] = \sigma^2$. $\leftarrow$ unbiased!

3) Tail bounds:

i) Markov's inequality: For any random variable $X \geq 0$ and any constant $a > 0$,
$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Pf: (Continuous case) $\mathbb{E}[X] = \int_0^\infty x\,\underset{\uparrow \text{ PDF of } X}{f_X(x)}\,dx = \underbrace{\int_0^a x f_X(x)\,dx}_{\geq 0} + \int_a^\infty x f_X(x)\,dx \geq \int_a^\infty x f_X(x)\,dx \geq a\int_a^\infty f_X(x)\,dx$

$$\Rightarrow \mathbb{E}[X] \geq a\,\mathbb{P}(X \geq a). \quad \blacksquare$$

ii) Chebyshev's inequality: For any random variable X and any constant $a>0$,
$$\mathbb{P}(|X-\mathbb{E}[X]| \geq a) \leq \frac{\text{var}(X)}{a^2}.$$

Pf: Consider the random variable $Y = (X-\mathbb{E}[X])^2 \geq 0$ with $\mathbb{E}[Y] = \text{var}(X)$. By Markov's inequality
$$\mathbb{P}(|X-\mathbb{E}[X]|^2 \geq a^2) = \mathbb{P}(Y \geq a^2) \leq \frac{\mathbb{E}[Y]}{a^2} = \frac{\text{var}(X)}{a^2}.$$

Since the event $\{|X-\mathbb{E}[X]| \geq a\}$ is equal to the event $\{Y \geq a^2\}$, we have
$$\mathbb{P}(|X-\mathbb{E}[X]| \geq a) \leq \frac{\text{var}(X)}{a^2}. \quad \blacksquare$$

ANURAN MAKUR

01/31/2023

4) Weak Law of Large Numbers (WLLN):

For i.i.d. random variables $X_1, \dots, X_n$ with mean $\mathbb{E}[X] = \mu$ and variance $\text{var}(X) = \sigma^2$, we have
$$\lim_{n \to \infty} \mathbb{P}\Big(\Big| \underbrace{\tfrac{1}{n}\sum_{i=1}^{n} X_i}_{\bar{X}} - \underbrace{\mu}_{\mathbb{E}[X]} \Big| \geq \varepsilon \Big) = 0 \quad \text{for all } \varepsilon > 0.$$

(Intuition: $\bar{X}$ is close to $\mathbb{E}[X]$ when n is large with high probability.)

Pf: For any $\varepsilon > 0$, by Chebyshev's inequality,
$$\mathbb{P}\Big(\Big| \underbrace{\tfrac{1}{n}\sum_{i=1}^{n} X_i}_{\bar{X}} - \underbrace{\mu}_{\mu = \mathbb{E}[\bar{X}]} \Big| \geq \varepsilon \Big) \leq \frac{\text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\varepsilon^2} = \frac{\text{var}(X)}{n\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Hence, $\lim_{n \to \infty} \mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} X_i - \mu \right| \geq \varepsilon \right) = 0$. ■

5) Central Limit Theorem (CLT):

Let $\Phi(x) \triangleq \int_{-\infty}^{x} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-t^2/2}}_{\text{normal PDF with mean 0 and variance 1}} dt$ be the standard normal CDF.

Let $X_1, \dots, X_n$ be i.i.d. random variables with mean $\mathbb{E}[X] = \mu$ and variance $\text{var}(X) = \sigma^2$.
Then, for all $x \in \mathbb{R}$, $\lim_{n \to \infty} \mathbb{P}\Big(\underbrace{\frac{1}{\sqrt{n}\,\sigma}\sum_{i=1}^{n}(X_i - \mu)}_{} \leq x \Big) = \Phi(x)$.

Intuition: CDF of this mean 0 and variance 1 random variable looks like normal CDF when n is large.

Anuran Makur

Some Probability Results on Induced Distributions

Date 02/14/2023 No.

① Change-of-Variables Formula: (linear case)

Let X be a continuous random variable with PDF f_X and CDF F_X. For any constants $a \neq 0$ and $b \in \mathbb{R}$, define the continuous random variable $Y = aX + b$. Suppose Y has CDF F_Y and PDF f_Y. Can we compute f_Y in terms of f_X?

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX + b \leq y) = \begin{cases} \mathbb{P}\left(X \leq \frac{y-b}{a}\right), & a > 0 \\ \mathbb{P}\left(X \geq \frac{y-b}{a}\right), & a < 0 \end{cases} = \begin{cases} F_X\left(\frac{y-b}{a}\right), & a > 0 \\ 1 - F_X\left(\frac{y-b}{a}\right), & a < 0 \end{cases}$$

$\uparrow$ $\mathbb{P}\left(X = \frac{y-b}{a}\right) = 0$ as X is continuous

$$\Rightarrow f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} \frac{d}{dy} F_X\left(\frac{y-b}{a}\right), & a > 0 \\ \frac{d}{dy}\left(1 - F_X\left(\frac{y-b}{a}\right)\right), & a < 0 \end{cases} \underset{\text{Chain rule}}{=} \begin{cases} f_X\left(\frac{y-b}{a}\right)\frac{1}{a}, & a > 0 \\ -f_X\left(\frac{y-b}{a}\right)\frac{1}{a}, & a < 0 \end{cases} = f_X\left(\frac{y-b}{a}\right)\frac{1}{|a|}$$

$\therefore$ For all $y \in \mathbb{R}$, $\boxed{f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)}$.

② Convolution: (discrete case)

Let X and Y be independent discrete random variables taking values in $\mathbb{Z}$ with PMFs f_X and f_Y, respectively. Define the discrete random variable $Z = X + Y$, which has PMF f_Z and also takes values in $\mathbb{Z}$. Can we compute f_Z in terms of f_X and f_Y?

$$\begin{aligned} f_Z(k) &= \mathbb{P}(Z = k) = \mathbb{P}(X + Y = k) = \mathbb{P}(\exists j \in \mathbb{Z}, X = j \text{ and } Y = k - j) \\ &= \sum_{j=-\infty}^{\infty} \mathbb{P}(X = j \text{ and } Y = k - j) \underset{\text{independence}}{=} \sum_{j=-\infty}^{\infty} \mathbb{P}(X = j)\,\mathbb{P}(Y = k - j) = \sum_{j=-\infty}^{\infty} f_X(j) f_Y(k - j) \end{aligned}$$

Define the convolution of f_X and f_Y as $(f_X \star f_Y)(k) \triangleq \sum_{j=-\infty}^{\infty} f_X(j) f_Y(k-j)$ for all $k \in \mathbb{Z}$.

$\therefore$ $\boxed{f_Z = f_X \star f_Y}$.

MATRIX CALCULUS & REGRESSION

Date 02/28/2023 No.

① Gradient:

For a differentiable function $f: \mathbb{R}^d \to \mathbb{R}$, $f(x_1, \dots, x_d)$, we define its gradient as the vector field $\nabla f: \mathbb{R}^d \to \mathbb{R}^d$ given by:

$$\forall x \in \mathbb{R}^d, \quad \nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_d}(x) \end{bmatrix}.$$

② Gradient of Quadratic Form:

Define the quadratic form $f(x) = x^T A x$ for any $x \in \mathbb{R}^d$, given a fixed matrix $A \in \mathbb{R}^{d \times d}$.

Prop: $\forall x \in \mathbb{R}^d$, $\nabla f(x) = (A + A^T)x$.

Pf: Observe that $f(x) = [x_1 \cdots x_d] \begin{bmatrix} [Ax]_1 \\ \vdots \\ [Ax]_d \end{bmatrix} = \sum_{i=1}^d x_i [Ax]_i = \sum_{i=1}^d x_i \sum_{j=1}^d A_{ij} x_j = \sum_{i=1}^d \sum_{j=1}^d x_i x_j A_{ij}$.

For any $k \in \{1, \dots, d\}$,

$$\frac{\partial f}{\partial x_k}(x) = \frac{\partial}{\partial x_k}\left(\sum_{i=1}^d \sum_{j=1}^d x_i x_j A_{ij}\right) = \frac{\partial}{\partial x_k}\left(x_k^2 A_{kk} + \sum_{i \neq k} x_i x_k A_{ik} + \sum_{j \neq k} x_k x_j A_{kj}\right)$$

$$= 2x_k A_{kk} + \sum_{i \neq k} x_i A_{ik} + \sum_{j \neq k} x_j A_{kj}$$

$$= \sum_{i=1}^d x_i \underbrace{A_{ik}}_{[A^T]_{ki}} + \sum_{j=1}^d x_j A_{kj}$$

$$= [A^T x]_k + [Ax]_k$$

$$= [(A + A^T)x]_k.$$

■

③ Gradient of Linear Form:

Define the linear form $f(x) = b^T x$ for any $x \in \mathbb{R}^d$, given a fixed vector $b \in \mathbb{R}^d$.

Prop: $\forall x \in \mathbb{R}^d$, $\nabla f(x) = b$.

Pf: For any $k \in \{1, \dots, d\}$, $\frac{\partial f}{\partial x_k}(x) = \frac{\partial}{\partial x_k}\left(\sum_{i=1}^d b_i x_i\right) = b_k$.

■

④ Normal Equation for Regression:

For any given target values $y \in \mathbb{R}^N$ & feature matrix $X \in \mathbb{R}^{N \times (M+1)}$, define the mean-squared error (MSE):

N data samples

no. of explanatory variables

$$\forall \beta \in \mathbb{R}^{M+1}, \quad E(\beta) \triangleq \frac{1}{N} \|y - X\beta\|_2^2 .$$

regression coefficients

To minimize $E(\beta)$ over β, we use the stationarity condition:

$$\nabla E(\beta) = 0 .$$

Hence, we have:

$$\begin{aligned} 0 &= \nabla\left(\frac{1}{N}\|y - X\beta\|_2^2\right) = \nabla\left(\frac{1}{N}(y - X\beta)^T(y - X\beta)\right) \\ &= \nabla\left(\frac{1}{N}(y^T - \beta^T X^T)(y - X\beta)\right) = \nabla\left(\frac{1}{N}\left(y^T y - 2(X^T y)^T \beta + \beta^T X^T X \beta\right)\right) \\ &= -\frac{2}{N} X^T y + \frac{1}{N}(X^T X + X^T X)\beta \quad \leftarrow \text{use gradient propositions} \end{aligned}$$

$$\Leftrightarrow \quad \boxed{X^T X \beta = X^T y} \Big\} \text{ NORMAL EQUATION}$$

ANURAN MAKUR

MEAN & VARIANCE OF REGRESSION COEFFICIENT

Date 03/02/2023 No.

① Setup:

Assume training samples $(x_1,y_1),\ldots,(\underset{\text{feature}}{x_N},\underset{\text{target/label}}{y_N}) \in \mathbb{R}^2$ are generated according to the simple linear model: $y_n = \alpha x_n + \beta + \varepsilon_n$, $n=1,\ldots,N$,
where $x_1,\ldots,x_n$ are fixed values, α & β are the true fixed parameters, and $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. noise random variables with mean 0 and variance σ^2.

Simple linear regression produces the estimate $a \triangleq \overbrace{\dfrac{\sum_{n=1}^{N} x_n y_n - N\bar{x}\bar{y}}{\sum_{n=1}^{N} x_n^2 - N\bar{x}^2}}^{\text{regression coefficient}}$ for α, where $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and $\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$.

② Mean of a:

Note that $\mathbb{E}[y_n] = \alpha x_n + \beta + \mathbb{E}[\varepsilon_n] = \alpha x_n + \beta$ and $\text{var}(y_n) = \text{var}(\varepsilon_n) = \sigma^2$ for $n=1,\ldots,N$.
Hence,

$$\mathbb{E}[a] = \mathbb{E}\left[\frac{\sum_{n=1}^{N}(x_n-\bar{x})y_n}{\sum_{n=1}^{N} x_n^2 - N\bar{x}^2}\right] = \frac{\sum_{n=1}^{N}(x_n-\bar{x})(\alpha x_n+\beta)}{\sum_{n=1}^{N} x_n^2 - N\bar{x}^2} = \frac{\alpha\sum_{n=1}^{N} x_n^2 - N\alpha\bar{x}^2}{\sum_{n=1}^{N} x_n^2 - N\bar{x}^2} + \frac{\cancel{\beta\left(\sum_{n=1}^{N} x_n - N\bar{x}\right)}^{\to 0}}{\sum_{n=1}^{N} x_n^2 - N\bar{x}^2}$$

$$\Rightarrow \boxed{\mathbb{E}[a] = \alpha}$$ ← a is an unbiased estimator of α

③ Variance of a:

Note that $\underline{\sum_{n=1}^{N}(x_n-\bar{x})^2} = \sum_{n=1}^{N} x_n^2 + N\bar{x}^2 - 2\bar{x}\sum_{n=1}^{N} x_n = \underline{\sum_{n=1}^{N} x_n^2 - N\bar{x}^2}$.

Hence,

$$\text{var}(a) = \text{var}\left(\frac{\sum_{n=1}^{N}(x_n-\bar{x})y_n}{\sum_{n=1}^{N}(x_n-\bar{x})^2}\right) \overset{\text{var additive if r.vs independent}}{=} \frac{1}{\left(\sum_{n=1}^{N}(x_n-\bar{x})^2\right)^2}\sum_{n=1}^{N}(x_n-\bar{x})^2 \underbrace{\text{var}(y_n)}_{=\sigma^2}$$

$$\Rightarrow \boxed{\text{var}(a) = \frac{\sigma^2}{\sum_{n=1}^{N}(x_n-\bar{x})^2}}$$

④ Standard Error when Testing whether $\alpha = 0$: (σ^2 unknown)

- Use $s^2 = \frac{1}{N-2}\sum_{n=1}^{N}(y_n-\hat{y}_n)^2$ to estimate σ^2, where $\hat{y}_n = a x_n + b$ ← regression intercept

  $y_n - \hat{y}_n = y_n - a x_n - b \approx y_n - \alpha x_n - \beta = \varepsilon_n \Rightarrow y_n - \hat{y}_n$ is an estimate of ε_n and $s^2 \approx \frac{1}{N}\sum_{n=1}^{N}(\varepsilon_n - 0)^2 \approx \sigma^2$
- $\mathbb{E}[s^2] = \sigma^2$ ← unbiased estimator of σ^2
- Use $\boxed{\underbrace{\widehat{\text{var}(a)}}_{(\text{standard error of } a)^2} = s^2 \Big/ \sum_{n=1}^{N}(x_n-\bar{x})^2}$ to estimate $\text{var}(a)$