Introduction to Data Science - 1MS041

Benny Avelin

Department of Mathematics

HT 2023

Recap

Common test metrics

- The mean squared error (MSE) (Usually used to measure model fit)

$$\mathbb{E}[(\hat{\phi}(X) - Y)^2 \mid \hat{\phi}]$$

and the root mean squared error (RMSE)

$$\sqrt{\mathbb{E}[(\hat{\phi}(X) - Y)^2 \mid \hat{\phi}]}.$$

- The mean absolute error (MAE)

$$\mathbb{E}[|\hat{\phi}(X) - Y| \mid \hat{\phi}]$$

often preferred as it is more explanatory.

- R^2, or explained variance

$$1 - \frac{\mathbb{E}[(\hat{\phi}(X) - Y)^2 \mid \hat{\phi}]}{\mathbb{V}(Y)}$$

Recap: Calibration

Calibration error

Consider f a given fixed function, then the calibration error is defined as

$$\sqrt{\mathbb{E}[\|\mathbb{E}[Y \mid f(X)] - f(X)\|^2]}$$

Note that

$$\mathbb{E}[\|Y - f(X)\|^2] = \mathbb{E}[\|\mathbb{E}[Y \mid f(X)] - f(X)\|^2] + \mathbb{E}[\|Y - \mathbb{E}[Y \mid f(X)]\|^2]$$

here we think about the first term as the bias^2 and the second term as variance. Thus we should interpret the calibration error as bias. The variance term should be considered as the variance of the prediction.

Calibration when we try to predict probabilities

Calibration error

Consider f a given fixed function predicting the probability of a label, then the calibration error is defined as

$$\sqrt{\mathbb{E}[\|\mathbb{P}[Y \mid f(X)] - f(X)\|^2]}$$

That is, the quantity $\mathbb{P}[Y \mid f(X)]$ is the true probability of the label when we predict the probability of the label being $f(X)$. Example, consider a model predicting $f(X) = 0.3$ for a group of samples, then $\mathbb{P}[Y \mid f(X) = 0.3]$ is the true probability of the label within those samples.

High dimension

Definition

Given a radius $r > 0$ we define the *d-dimensional ball* as the set

$$B_r(x) := \{y \in \mathbb{R}^d : |x - y| < r\}.$$

We also denote the *d-dimensional sphere* as the set

$$S_r(x) := \{y \in \mathbb{R}^d : |x - y| = r\}.$$

Whenever $r = 1$ we call $B_1(x), S_1(x)$ *unit ball* and *unit sphere* respectively. If $x = 0$ we omit it from the notation, and use $B_r = B_r(0)$ and $S_r = S_r(0)$.

Simulation using the normalized Gaussian

Definition

Let Y be a random variable taking values in $\mathbb{R}^d$ with density

$$f(x) = \exp\left(-\pi|x|^2\right), \quad x \in \mathbb{R}^d,$$

then it is called a normalized Gaussian.

Simulation using the normalized Gaussian

Definition

Let Y be a random variable taking values in $\mathbb{R}^d$ with density

$$f(x) = \exp\left(-\pi|x|^2\right), \quad x \in \mathbb{R}^d,$$

then it is called a normalized Gaussian.

Probability of landing inside a cube

Consider the unit ball B_1, and consider the probability

$$\mathbb{P}(Y \in B_1) = \int_{B_1} \exp\left(-\pi|x|^2\right) dx \geq \frac{|B_1|}{e^{\pi}}$$

Scaling of dimension

Lemma

Let $E \subset \mathbb{R}^d$ *and let* $\epsilon \in (0, 1]$, *then*

$$(1 - \epsilon)^d |E| = |(1 - \epsilon)E|$$

where $(1 - \epsilon)E := \{(1 - \epsilon)x : x \in E\}$.

Volume of the unit ball

Theorem

The volume of the unit ball in d dimensions is

$$|B_1| = \frac{2\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}$$

where Γ is the Gamma-function. For even dimensions we get

$$|B_1| = \frac{2\pi^{\frac{d}{2}}}{d(\frac{d}{2}-1)!}.$$

Volume of the unit ball

Theorem

The volume of the unit ball in d dimensions is

$$|B_1| = \frac{2\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}$$

where Γ is the Gamma-function. For even dimensions we get

$$|B_1| = \frac{2\pi^{\frac{d}{2}}}{d(\frac{d}{2}-1)!}.$$

Question

Let us assume we want to produce a sample from the uniform distribution in the unit ball using rejection sampling and using the uniform distribution on the unit cube as sampling distribution, what happens?

Unit sphere

Model

We say that a $\mathbb{R}^d$ valued random variable Z is *uniform at random from the unit sphere* if $Z \in S_1$ and for any A we have

$$\mathbb{P}(Z \in A) = \frac{1}{|S_1|} \int_{S_1} \mathbb{1}_A(\theta) d\Omega(\theta)$$

where the integral above is the surface integral on the sphere, here $d\Omega$ is the surface element on S_1. We denote this as $Z \sim \text{uniform}(S_1)$.

Unit sphere

Model

We say that a $\mathbb{R}^d$ valued random variable Z is *uniform at random from the unit sphere* if $Z \in S_1$ and for any A we have

$$\mathbb{P}(Z \in A) = \frac{1}{|S_1|} \int_{S_1} \mathbb{1}_A(\theta) d\Omega(\theta)$$

where the integral above is the surface integral on the sphere, here $d\Omega$ is the surface element on S_1. We denote this as $Z \sim \text{uniform}(S_1)$.

Gaussian trick

If we consider Z coming from a "spherical Gaussian", then

$$\frac{Z}{|Z|} \sim \text{uniform}(S_1).$$

Unit ball

Model

We say that an $\mathbb{R}^d$ valued random variable Z is *uniform at random from the unit ball* if $Z \in B_1$ and for any A we have

$$\mathbb{P}(Z \in A) = \frac{1}{|B_1|} \int_{B_1} \mathbb{1}_A(z) dz = \frac{|A \cap B_1|}{|B_1|}.$$

In short, the probability of landing inside $A \cap B_1$ is given by the proportion of the volume it makes up out of B_1. We say $Z \sim \text{uniform}(B_1)$.

The annulus theorem

Model

A continuous $\mathbb{R}^d$ valued random variable Z with density function

$$f(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}|x|^2\right), \quad x \in \mathbb{R}^d$$

is called a *spherical Gaussian*.

The annulus theorem

Model

A continuous $\mathbb{R}^d$ valued random variable Z with density function

$$f(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}|x|^2\right), \quad x \in \mathbb{R}^d$$

is called a *spherical Gaussian*.

Note

For a spherical Gaussian Z

$$\mathbb{E}[|Z|^2] = d.$$

The annulus theorem

Theorem

For a d-dimensional spherical Gaussian Z, then for any $\beta \leq \sqrt{d}$ we have

$$\mathbb{P}\left(\sqrt{d} - \beta \leq |X| \leq \sqrt{d} + \beta\right) < 2e^{-\frac{\beta^2}{128}}.$$

Johnson-Lindenstrauss lemma

Theorem (Projection)

Let v be a fixed vector in $\mathbb{R}^d$ of length 1, fix $\epsilon \in (0,1)$ and let $U_1, \ldots, U_k \in \mathbb{R}^d$ be a spherical Gaussian. Consider the projection onto $(U_1, \ldots, U_k)$

$$f(v) = (U_1 \cdot v, \ldots, U_k \cdot v) : \mathbb{R}^d \to \mathbb{R}^k,$$

then

$$\mathbb{P}\left(\left| |f(v)| - \sqrt{k}|v| \right| \geq \epsilon\sqrt{k}|v| \right) \leq 2e^{-\frac{k\epsilon^2}{128}}.$$

Johnson-Lindenstrauss lemma

Theorem (Johnson-Lindenstrauss)

For any $0 < \epsilon < 1$ *and any integer* n*, let* $k > \frac{384 \ln(n)}{\epsilon^2}$*. For any set of* n *points* $\{v_1, \dots, v_n\} \in \mathbb{R}^d$ *then the random projection defined previously satisfies*

$$\mathbb{P}\left((1-\epsilon)\sqrt{k}|v_i - v_j| \leq |f(v_i - v_j)| \leq (1+\epsilon)\sqrt{k}|v_i - v_j|\right) \geq 1 - \frac{3}{2n}$$