Introduction to Data Science - 1MS041

Benny Avelin

Department of Mathematics

HT 2023

Recall from last time

- We saw an example of different ways to construct estimators for a problem, and we calculated their standard errors. All estimators are not created equal.

Recall from last time

- We saw an example of different ways to construct estimators for a problem, and we calculated their standard errors. All estimators are not created equal.
- We explored the log-Loss, i.e. $L(z, \alpha) = -\ln p_\alpha(z)$, where p_α is a proposal density for our data, we assume that there is an α^* such that the data comes from p_{α^*}.

Recall from last time

- We saw an example of different ways to construct estimators for a problem, and we calculated their standard errors. All estimators are not created equal.
- We explored the log-Loss, i.e. $L(z, \alpha) = -\ln p_\alpha(z)$, where p_α is a proposal density for our data, we assume that there is an α^* such that the data comes from p_{α^*}.
- We saw that the empirical risk is the negative log Likelihood

$$\hat{R}(\alpha) := \frac{1}{n}\sum_{i=1}^{n}(-\ln(p_\alpha(X_i)))$$

$$R(\alpha) = \mathbb{E}[-\ln(p_\alpha(X))]$$

Recall from last time

- We explored the problem of estimating the σ in $N(0, \sigma^2)$ using the Likelihood.

Recall from last time

- We explored the problem of estimating the σ in $N(0, \sigma^2)$ using the Likelihood.
- We considered the conditional likelihood, i.e. our proposal density is of the form $f_\alpha(x, y) = p_\alpha(y \mid x)p(x)$ for some fixed $p(x)$.

Recall from last time

- We explored the problem of estimating the σ in $N(0, \sigma^2)$ using the Likelihood.
- We considered the conditional likelihood, i.e. our proposal density is of the form $f_\alpha(x, y) = p_\alpha(y \mid x) p(x)$ for some fixed $p(x)$.
- We saw

Recall from last time

- We explored the problem of estimating the σ in $N(0, \sigma^2)$ using the Likelihood.
- We considered the conditional likelihood, i.e. our proposal density is of the form $f_\alpha(x, y) = p_\alpha(y \mid x)p(x)$ for some fixed $p(x)$.
- We saw
    - $p_{\alpha^*, X} = N(\alpha_1 X + \alpha_2, \alpha_3^2)$, Linear regression

Recall from last time

- We explored the problem of estimating the σ in $N(0, \sigma^2)$ using the Likelihood.
- We considered the conditional likelihood, i.e. our proposal density is of the form $f_\alpha(x, y) = p_\alpha(y \mid x)p(x)$ for some fixed $p(x)$.
- We saw
  - $p_{\alpha^*, X} = N(\alpha_1 X + \alpha_2, \alpha_3^2)$, Linear regression
  - $p_{\alpha^*, X} = \text{Bernoulli}(G(\alpha_1 X + \alpha_2))$,

$$G(x) = \frac{1}{1 + e^{-x}}$$

    Logistic regression

Today

We will explore how to generate random variables on the computer and what that really means.

Today

We will explore how to generate random variables on the computer and what that really means.

Definition (Informal)

A **uniform pseudorandom number generator** (UPRNG) is an algorithm which starting from an initial value u_0 and a transformation D, produces a sequence $u_i = D(u_{i-1})$ in $[0,1]$ for $i = 1, \ldots$. For all n, $u_1, \ldots, u_n$ approximate the behavior of an i.i.d. sequence of uniform($[0,1]$) random numbers.

Pseudorandom

Definition (pseudorandom)

Consider the finite set $\mathcal{M} = \{0, 1, \ldots, M-1\}$ and consider the sequence $u_0, u_1, \ldots \in \mathcal{M}$. For every $a \in \mathcal{M}$, define $N_n(a)$ as the number of $u_i = a$ for $i = 0, 1, 2, \ldots, n-1$. We call the sequence $u_0, u_1, \ldots$ **pseudorandom** on $\mathcal{M}$ if and only if for every $a \in \mathcal{M}$

$$\frac{N_n(a)}{n} \to \frac{1}{M}.$$

Congruential generators

Definition

Let u_0 be fixed and let D be a map, define the dynamical system

$$u_i = D(u_{i-1}), \quad i = 1, \ldots$$

We call T_0 the period of D started at u_0 the smallest positive integer such that

$$u_{i+T_0} = u_i, \text{ for some } i.$$

The smallest period T for all admissible starting points u_0 is called the period for D.

Definition

A **congruential generator** with parameters (a, b, M) on $\{0, 1, \dots, M-1\}$ is defined by the function

$$D(x) = (ax + b) \mod M.$$

Full period

The following number theoretical theorem tells us exactly when we can expect period M.

Theorem (Hull–Dobell Theorem)

The congruential generator (a, b, M) *has period* M *iff*

- $\gcd(b, M) = 1$,
- p *divides* $a - 1$ *for every prime* p *that divides* M
- 4 *divides* $a - 1$ *if* 4 *divides* M.

Remark

Consider a congruential generator D on $\mathcal{M} = \{0, 1, \ldots, M-1\}$ with period M, then for any starting point $u_0 \in \mathcal{M}$, the sequence $u_i = D(u_{i-1})$ is pseudorandom on $\mathcal{M}$.

Remark

Consider a congruential generator D on $\mathcal{M} = \{0, 1, \ldots, M-1\}$ with period M, then for any starting point $u_0 \in \mathcal{M}$, the sequence $u_i = D(u_{i-1})$ is pseudorandom on $\mathcal{M}$.

Lemma

Consider a congruential generator D on $\mathcal{M} = \{0, 1, \ldots, M-1\}$ with period M that is divisible by K, then for any starting point $u_0 \in \mathcal{M}$, define $u_i = D(u_{i-1})$ then the sequence $v_i = \lfloor (u_i/M) * K \rfloor$ for $1 \leq K \leq M$ is pseudorandom on $\mathcal{K} = \{0, 1, \ldots, K-1\}$ if M is a multiple of K.

Note

If we define the map $D'(a, b) = (\lfloor (D(b)/M) * K \rfloor, D(b))$, then the period of D' is M.

Prototype

If we instead consider $v_i = u_i/M$ we will get numbers between 0 and 1, and we have a prototype of a **uniform pseudorandom number generator.**

Conclusion

1. Find a congruential generator with large period

Conclusion

1. Find a congruential generator with large period
2. If we want to produce uniform distribution over $0, 1, \ldots, K$ we just divide by the period and multiply by K.

Conclusion

1. Find a congruential generator with large period
2. If we want to produce uniform distribution over $0, 1, \ldots, K$ we just divide by the period and multiply by K.
3. If we want to produce uniform numbers between 0 and 1 we instead just divide by the period.

Getting to the uniform$[0,1]$

Lemma

Let $u_0, u_1, \ldots$ be a psuedo random sequence over $\mathcal{M} = \{0, 1, \ldots, M-1\}$. Then $v_i = u_i/M$ has the empirical mean and variance limits as follows

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} v_i = \frac{1}{2} - \frac{1}{2M}$$

and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} v_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} v_i\right)^2 = \frac{1}{12} - \frac{1}{12M^2}.$$

Uniform Pseudo Random Generator

Does this now give us a uniform pseudo random generator?

Uniform Pseudo Random Generator

Does this now give us a uniform pseudo random generator?

Lemma

Let $v_0, v_1, \ldots$ be a pseudorandom sequence in $\mathcal{M} = \{0, 1, \ldots, M-1\}$, define $u_i = v_i/M$. For any interval $A = (a, b) \subset [0, 1]$, define $N_n(A)$ as the number of $u_i \in A$ for $i = 0, 1, 2, \ldots, n-1$. We have

$$\left| \lim_{n\to\infty} \frac{N_n(A)}{n} - \int_A dx \right| \leq \frac{1}{M}.$$