

Natural Language Models and Interfaces

BSc Artificial Intelligence

Lecturer: Wilker Aziz

Institute for Logic, Language, and Computation

2018, week 1, lecture b

Random variables

Probability distributions

Discrete distributions

Maximum likelihood estimation

Variables: Deterministic vs Random

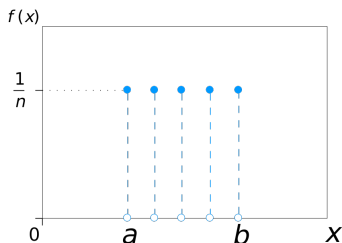
Deterministic variable: $v = 5$

Image from Wikipedia

Variables: Deterministic vs Random

Deterministic variable: $v = 5$

Random variable: $X \sim \mathcal{U}(a, b)$



- ▶ the random variable can take on **any value** in a certain set
- ▶ here this set is the discrete interval $[a, b]$
- ▶ we don't know the **value** of the random variable
we know it's **distribution**

Probability of an outcome

We cannot talk about **the exact value** of the random variable but we can reason about its **possible values**

- ▶ we quantify the **degree of belief** we have in each **outcome**

Uniform distribution: every outcome is **equally likely**

- ▶ if n is the size of the set of possible outcomes the **probability** that X takes on any value (e.g. a) is $\frac{1}{n}$
 $P(X = x) = \frac{1}{n}$ for all $x \in [a, b]$

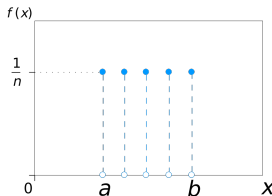


Image from Wikipedia

Let's name some things

A random variable is a **function**

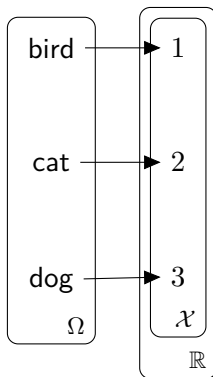
- ▶ it maps from a **sample space** Ω to \mathbb{R}
 $X : \Omega \rightarrow \mathbb{R}$

Example: “which pet do kids love the most?”

- ▶ Sample space: $\Omega = \{\text{bird}, \text{cat}, \text{dog}\}$

$$X(\omega) = \begin{cases} 1 & \omega = \{\text{bird}\} \\ 2 & \omega = \{\text{cat}\} \\ 3 & \omega = \{\text{dog}\} \end{cases}$$

- ▶ if say $X = x$ we mean the set of outcomes $\{\omega : X(\omega) = x\}$ which is called an **event**
- ▶ we call \mathcal{X} the **support** of X



Temperature example

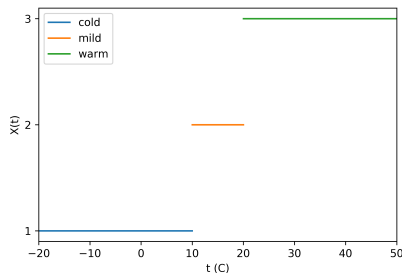
Let's take the outside temperature as a random variable

- ▶ we might not particularly care whether it's -3 or -3.2
- ▶ but we probably care to ask
"How does it feel outside?"

Let's define an RV

- ▶ Sample space
some segment of the real line
 - ▶ perhaps from -40 to 50 ?
 - ▶ cap on precision?

- ▶
$$X(t) = \begin{cases} 1 & t < 10 \\ 2 & 10 \leq t \leq 20 \\ 3 & t > 20 \end{cases}$$



Types of random variables

Random variables are different in nature

- ▶ categorical: toss a coin
- ▶ ordinal: number of items in a bag
- ▶ continuous: height, weight

They can have finite or infinite support

- ▶ toss a coin, throw a die: finitely many outcomes
- ▶ distances: infinitely many outcomes
- ▶ number of stars: infinitely many outcomes

They can be vector-valued

- ▶ a point in a 2D-plane: e.g. (x, y) coordinates
- ▶ a point in a d -dimensional space: e.g. database records
house: *floor area, latitude, longitude, altitude, number of rooms, age, number of past owners, market value*

NLMI

Random variables

Probability distributions

Discrete distributions

Maximum likelihood estimation

Discrete probability distribution

The discrete probability distribution of a random variable X

- ▶ assigns a **probability value** to each value X may take on
- ▶ probability values are *never less than 0*
 $P(X = x) \geq 0$ for all $x \in \mathcal{X}$
- ▶ and a probability distribution *sums to 1*
 $\sum_{x \in \mathcal{X}} P(X = x) = 1$
- ▶ thus we have
 - ▶ $0 \leq P(X = x) \leq 1$ for all $x \in \mathcal{X}$
 - ▶ $P(X \neq x) = 1 - P(X = x)$

Discrete probability distribution

The discrete probability distribution of a random variable X

- ▶ assigns a **probability value** to each value X may take on
- ▶ probability values are *never less than 0*
 $P(X = x) \geq 0$ for all $x \in \mathcal{X}$
- ▶ and a probability distribution *sums to 1*
 $\sum_{x \in \mathcal{X}} P(X = x) = 1$
- ▶ thus we have
 - ▶ $0 \leq P(X = x) \leq 1$ for all $x \in \mathcal{X}$
 - ▶ $P(X \neq x) = 1 - P(X = x)$

Notation

- ▶ distribution: $P_X, P_X(X), P(X)$
- ▶ value: $P_X(X = x), P(X = x), P_X(x), P(x)$

Joint probability distribution

Oftentimes we care about multiple random variables
and how **their outcomes co-occur**

| Ω | | Letter (L) | | P_{GL} | | Letter (L) | |
|-----------|-----|----------------|--------|-----------|-----|----------------|------|
| Grade | G | 0 | 1 | Grade | G | 0 | 1 |
| $[0, 6)$ | 1 | (1, 0) | (1, 1) | $[0, 6)$ | 1 | 0.16 | 0.04 |
| $[6, 8)$ | 2 | (2, 0) | (2, 1) | $[6, 8)$ | 2 | 0.42 | 0.28 |
| $[8, 10]$ | 3 | (3, 0) | (3, 1) | $[8, 10]$ | 3 | 0.01 | 0.09 |

Table: Joint sample space Ω and joint distribution P_{GL}

Joint probability distribution

Oftentimes we care about multiple random variables
and how **their outcomes co-occur**

| Ω | | Letter (L) | | P_{GL} | | Letter (L) | |
|-----------|-----|----------------|--------|-----------|-----|----------------|------|
| Grade | G | 0 | 1 | Grade | G | 0 | 1 |
| $[0, 6)$ | 1 | (1, 0) | (1, 1) | $[0, 6)$ | 1 | 0.16 | 0.04 |
| $[6, 8)$ | 2 | (2, 0) | (2, 1) | $[6, 8)$ | 2 | 0.42 | 0.28 |
| $[8, 10]$ | 3 | (3, 0) | (3, 1) | $[8, 10]$ | 3 | 0.01 | 0.09 |

Table: Joint sample space Ω and joint distribution P_{GL}

Joint probability $P(G = g, L = l)$

- ▶ we refer to the event $\{\omega : G(\omega) = g, L(\omega) = l\}$

Joint probability distribution

Oftentimes we care about multiple random variables
and how **their outcomes co-occur**

| Ω | | Letter (L) | | P_{GL} | | Letter (L) | |
|-----------|-----|----------------|--------|-----------|-----|----------------|------|
| Grade | G | 0 | 1 | Grade | G | 0 | 1 |
| $[0, 6)$ | 1 | (1, 0) | (1, 1) | $[0, 6)$ | 1 | 0.16 | 0.04 |
| $[6, 8)$ | 2 | (2, 0) | (2, 1) | $[6, 8)$ | 2 | 0.42 | 0.28 |
| $[8, 10]$ | 3 | (3, 0) | (3, 1) | $[8, 10]$ | 3 | 0.01 | 0.09 |

Table: Joint sample space Ω and joint distribution P_{GL}

Joint probability $P(G = g, L = l)$

- ▶ we refer to the event $\{\omega : G(\omega) = g, L(\omega) = l\}$

Properties

- ▶ $0 \leq P(G = g, L = l) \leq 1$ for all $(g, l) \in \mathcal{G} \times \mathcal{L}$
- ▶ $\sum_{g \in \mathcal{G}} \sum_{l \in \mathcal{L}} P(G = g, L = l) = 1$

Marginal probability

Recover the distribution of each RV

| P_{GL} | | Letter (L) | | P_G |
|-----------|-----|----------------|------|-------|
| Grade | G | 0 | 1 | |
| $[0, 6)$ | 1 | 0.16 | 0.04 | 0.2 |
| $[6, 8)$ | 2 | 0.42 | 0.28 | 0.7 |
| $[8, 10]$ | 3 | 0.01 | 0.09 | 0.1 |
| P_L | | 0.59 | 0.41 | |

Table: Joint distribution P_{GL} and marginals P_G and P_L

Sum over all values of one of the RVs

- ▶ $P(G = g) = \sum_{l \in \mathcal{L}} P(G = g, L = l)$
- ▶ $P(L = l) = \sum_{g \in \mathcal{G}} P(G = g, L = l)$

Conditional probability

If we know the value of one of the RVs
we can rescale to get a distribution

| P_{GL} | | Letter (L) | | P_G |
|-----------|-----|----------------|------|-------|
| Grade | G | 0 | 1 | |
| $[0, 6)$ | 1 | 0.16 | 0.04 | 0.2 |
| $[6, 8)$ | 2 | 0.42 | 0.28 | 0.7 |
| $[8, 10]$ | 3 | 0.01 | 0.09 | 0.1 |
| P_L | | 0.59 | 0.41 | |

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

| $P_{L G=g}$ | | Letter (L) | | | $P_{G L=l}$ | | Letter (L) | |
|--------------|-----|----------------|-----|---------------|-------------|-----|----------------|------|
| Grade | G | 0 | 1 | \rightarrow | Grade | G | 0 | 1 |
| $[0, 6)$ | 1 | 0.8 | 0.2 | 1.0 | $[0, 6)$ | 1 | 0.27 | 0.10 |
| $[6, 8)$ | 2 | 0.6 | 0.4 | 1.0 | $[6, 8)$ | 2 | 0.71 | 0.68 |
| $[8, 10]$ | 3 | 0.1 | 0.9 | 1.0 | $[8, 10]$ | 3 | 0.02 | 0.22 |
| \downarrow | | | | | | | 1.00 | 1.00 |

Table: Conditional distributions $P_{L|G=g}$ and $P_{G|L=l}$

Rules of probability

Chain rule

- ▶ Two RVs

$$P(X = x, Y = y) = P(X = x)P(Y = y|X = x)$$

Rules of probability

Chain rule

- ▶ Two RVs

$$P(X = x, Y = y) = P(X = x)P(Y = y|X = x)$$

- ▶ General ($n > 2$)

$$P(x_1, \dots, x_n) = P(x_1) \prod_{i=2}^n P(x_i|x_1, \dots, x_{i-1})$$

Rules of probability

Chain rule

- ▶ Two RVs

$$P(X = x, Y = y) = P(X = x)P(Y = y|X = x)$$

- ▶ General ($n > 2$)

$$P(x_1, \dots, x_n) = P(x_1) \prod_{i=2}^n P(x_i|x_1, \dots, x_{i-1})$$

Bayes rule

- ▶ if we know P_X and $P_{Y|X}$, we know the joint P_{XY}

Rules of probability

Chain rule

- ▶ Two RVs

$$P(X = x, Y = y) = P(X = x)P(Y = y|X = x)$$

- ▶ General ($n > 2$)

$$P(x_1, \dots, x_n) = P(x_1) \prod_{i=2}^n P(x_i|x_1, \dots, x_{i-1})$$

Bayes rule

- ▶ if we know P_X and $P_{Y|X}$, we know the joint P_{XY}
- ▶ then we can infer P_Y by marginalisation

Rules of probability

Chain rule

- ▶ Two RVs

$$P(X = x, Y = y) = P(X = x)P(Y = y|X = x)$$

- ▶ General ($n > 2$)

$$P(x_1, \dots, x_n) = P(x_1) \prod_{i=2}^n P(x_i|x_1, \dots, x_{i-1})$$

Bayes rule

- ▶ if we know P_X and $P_{Y|X}$, we know the joint P_{XY}
- ▶ then we can infer P_Y by marginalisation
- ▶ then we can infer $P_{X|Y}$

Rules of probability

Chain rule

- ▶ Two RVs

$$P(X = x, Y = y) = P(X = x)P(Y = y|X = x)$$

- ▶ General ($n > 2$)

$$P(x_1, \dots, x_n) = P(x_1) \prod_{i=2}^n P(x_i|x_1, \dots, x_{i-1})$$

Bayes rule

- ▶ if we know P_X and $P_{Y|X}$, we know the joint P_{XY}
- ▶ then we can infer P_Y by marginalisation
- ▶ then we can infer $P_{X|Y}$

$$P_{X|Y}(x|y) = \frac{P_X(x)P_{Y|X}(y|x)}{P_Y(y)}$$

Independence

If X does not depend on Y

we say X is independent of Y or $X \perp Y$

it holds that $P_{X|Y}(x|y) = P_X(x)$

Independence

If X does not depend on Y

we say X is independent of Y or $X \perp Y$

it holds that $P_{X|Y}(x|y) = P_X(x)$

This implies that for $X \perp Y$

$$P_{XY}(x, y) = P_X(x)P_Y(y)$$

Independence

If X does not depend on Y

we say X is independent of Y or $X \perp Y$

it holds that $P_{X|Y}(x|y) = P_X(x)$

This implies that for $X \perp Y$

$$P_{XY}(x, y) = P_X(x)P_Y(y)$$

And in general

$$P_{X_1^n}(x_1, \dots, x_n) = \prod_{i=1}^n P_{X_i}(x_i)$$

NLMI

Random variables

Probability distributions

Discrete distributions

Maximum likelihood estimation

Bernoulli

A **Bernoulli** variable is a binary random variable

$$X \sim \text{Bern}(p)$$

- ▶ $\mathcal{X} = \{0, 1\}$
- ▶ p is the **Bernoulli parameter**
 $0 \leq p \leq 1$
- ▶ $P(X = 1) = p$
- ▶ $P(X = 0) =$

Bernoulli

A **Bernoulli** variable is a binary random variable

$$X \sim \text{Bern}(p)$$

- ▶ $\mathcal{X} = \{0, 1\}$
- ▶ p is the **Bernoulli parameter**
 $0 \leq p \leq 1$
- ▶ $P(X = 1) = p$
- ▶ $P(X = 0) = 1 - p$

▶ Quiz

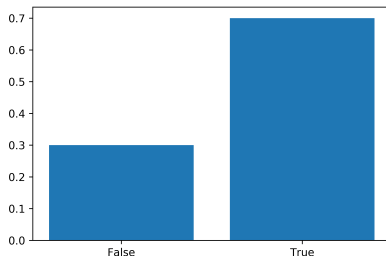
Bernoulli

A **Bernoulli** variable is a binary random variable

$$X \sim \text{Bern}(p)$$

- ▶ $\mathcal{X} = \{0, 1\}$
- ▶ p is the **Bernoulli parameter**
 $0 \leq p \leq 1$
- ▶ $P(X = 1) = p$
- ▶ $P(X = 0) = 1 - p$

▶ Quiz



Categorical

A **Categorical** variable can model 1 of k categories

$$X \sim \text{Cat}(\theta_1, \dots, \theta_k)$$

- ▶ $\mathcal{X} = \{1, \dots, k\}$
- ▶ the categorical parameter is a **probability vector**
 - ▶ $0 \leq \theta_x \leq 1$ for $x \in [1, k]$
 - ▶ $\sum_{x=1}^k \theta_x = 1$
- ▶ $P(X = x) = \theta_x$

Categorical

A **Categorical** variable can model 1 of k categories

$$X \sim \text{Cat}(\theta_1, \dots, \theta_k)$$

- ▶ $\mathcal{X} = \{1, \dots, k\}$
- ▶ the categorical parameter is a **probability vector**
 - ▶ $0 \leq \theta_x \leq 1$ for $x \in [1, k]$
 - ▶ $\sum_{x=1}^k \theta_x = 1$
- ▶ $P(X = x) = \theta_x$

▶ Quiz

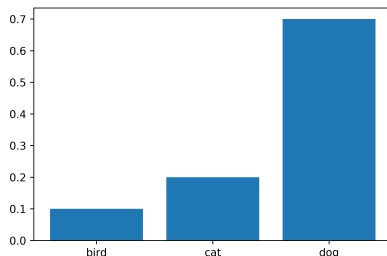
Categorical

A **Categorical** variable can model 1 of k categories

$$X \sim \text{Cat}(\theta_1, \dots, \theta_k)$$

- ▶ $\mathcal{X} = \{1, \dots, k\}$
- ▶ the categorical parameter is a **probability vector**
 - ▶ $0 \leq \theta_x \leq 1$ for $x \in [1, k]$
 - ▶ $\sum_{x=1}^k \theta_x = 1$
- ▶ $P(X = x) = \theta_x$

▶ Quiz



Random variables

Probability distributions

Discrete distributions

Maximum likelihood estimation

Statistical estimation

We investigate problems

- ▶ we hypothesise interactions between variables
- ▶ we assume variables have a certain nature
- ▶ we choose probability distributions
- ▶ we try to estimate parameters for these distributions as to reproduce “natural” observations

Likelihood

Let's imagine we are interested in a random phenomenon

- ▶ which we express with an rv $X \sim P_X$

Likelihood

Let's imagine we are interested in a random phenomenon

- ▶ which we express with an rv $X \sim P_X$

Then suppose we observe n realisations of the rv

- ▶ **observations** x_1, \dots, x_n
 $X_i \sim P_X$ for $i = 1, \dots, n$

Likelihood

Let's imagine we are interested in a random phenomenon

- ▶ which we express with an rv $X \sim P_X$

Then suppose we observe n realisations of the rv

- ▶ **observations** x_1, \dots, x_n
 $X_i \sim P_X$ for $i = 1, \dots, n$
- ▶ assume these observations are the result of **independent trials** (i.e. independent repetitions) of **the same random experiment**

Likelihood

Let's imagine we are interested in a random phenomenon

- ▶ which we express with an rv $X \sim P_X$

Then suppose we observe n realisations of the rv

- ▶ **observations** x_1, \dots, x_n
 $X_i \sim P_X$ for $i = 1, \dots, n$
- ▶ assume these observations are the result of **independent trials** (i.e. independent repetitions) of **the same random experiment**
- ▶ we call them *independent and identically distributed* observations

Likelihood

Let's imagine we are interested in a random phenomenon

- ▶ which we express with an rv $X \sim P_X$

Then suppose we observe n realisations of the rv

- ▶ **observations** x_1, \dots, x_n
 $X_i \sim P_X$ for $i = 1, \dots, n$
- ▶ assume these observations are the result of **independent trials** (i.e. independent repetitions) of **the same random experiment**
- ▶ we call them *independent and identically distributed* observations

From *independence* we know that $P_{X_1^n}(x_1, \dots, x_n) = \prod_{i=1}^n P_{X_i}(x_i)$

Likelihood

Let's imagine we are interested in a random phenomenon

- ▶ which we express with an rv $X \sim P_X$

Then suppose we observe n realisations of the rv

- ▶ **observations** x_1, \dots, x_n
 $X_i \sim P_X$ for $i = 1, \dots, n$
- ▶ assume these observations are the result of **independent trials** (i.e. independent repetitions) of **the same random experiment**
- ▶ we call them ***independent and identically distributed*** observations

From ***independence*** we know that $P_{X_1^n}(x_1, \dots, x_n) = \prod_{i=1}^n P_{X_i}(x_i)$
and with ***iid*** observations $\prod_{i=1}^n P_{X_i}(x_i) = \prod_{i=1}^n P_X(x_i)$

Parametric model

The problem is that we do not know P_X

Parametric model

The problem is that we do not know P_X

- ▶ but we can pick a family that makes sense

Parametric model

The problem is that we do not know P_X

- ▶ but we can pick a family that makes sense
 - ▶ $P_X := \text{Bern}(p)$ for coins
 - ▶ $P_X := \text{Cat}(\theta_1, \dots, \theta_k)$ for pet preferences

Parametric model

The problem is that we do not know P_X

- ▶ but we can pick a family that makes sense
 - ▶ $P_X := \text{Bern}(p)$ for coins
 - ▶ $P_X := \text{Cat}(\theta_1, \dots, \theta_k)$ for pet preferences
- ▶ for a fixed family, each choice of parameter gives us a new distribution

Parametric model

The problem is that we do not know P_X

- ▶ but we can pick a family that makes sense
 - ▶ $P_X := \text{Bern}(p)$ for coins
 - ▶ $P_X := \text{Cat}(\theta_1, \dots, \theta_k)$ for pet preferences
- ▶ for a fixed family, each choice of parameter gives us a new distribution
- ▶ we write $P_X(X; \alpha)$, or $P_{X;\alpha}$
to stress the dependency on a collection of parameters α

Parametric model

The problem is that we do not know P_X

- ▶ but we can pick a family that makes sense
 - ▶ $P_X := \text{Bern}(p)$ for coins
 - ▶ $P_X := \text{Cat}(\theta_1, \dots, \theta_k)$ for pet preferences
- ▶ for a fixed family, each choice of parameter gives us a new distribution
- ▶ we write $P_X(X; \alpha)$, or $P_{X;\alpha}$
to stress the dependency on a collection of parameters α

The maximum likelihood principle is about

- ▶ picking α to give maximum probability to observations

Parametric model

The problem is that we do not know P_X

- ▶ but we can pick a **family** that makes sense
 - ▶ $P_X := \text{Bern}(p)$ for coins
 - ▶ $P_X := \text{Cat}(\theta_1, \dots, \theta_k)$ for pet preferences
- ▶ for a fixed family, each choice of parameter gives us a new distribution
- ▶ we write $P_X(X; \alpha)$, or $P_{X;\alpha}$ to stress the dependency on a collection of parameters α

The **maximum likelihood principle** is about

- ▶ picking α to give maximum probability to observations
- ▶ where the probability of observations (or **likelihood**) is
$$P(x_1, \dots, x_n; \alpha) = \prod_{i=1}^n P_{X;\alpha}(x_i)$$
due to the **idd** assumption

Optimisation

We start with our **likelihood function**

$$P(x_1, \dots, x_n; \alpha) = \prod_{i=1}^n P_{X; \alpha}(x_i)$$

which depends on a choice of α

And we proceed by optimising this choice

$$\alpha^* = \underset{\alpha}{\operatorname{argmax}} P(x_1, \dots, x_n; \alpha) \quad \alpha \text{ such that likelihood is maximised}$$

$$= \underset{\alpha}{\operatorname{argmax}} \prod_{i=1}^n P_X(x_i; \alpha) \quad \text{iid observations}$$

$$= \underset{\alpha}{\operatorname{argmax}} \log \prod_{i=1}^n P_X(x_i; \alpha) \quad \log \text{ is } \text{monotonic}$$

$$= \underset{\alpha}{\operatorname{argmax}} \sum_{i=1}^n \log P_X(x_i; \alpha) \quad \text{numerically convenient}$$

We assume argmax to return a point (not a set). Want to know more about argmax ? Check [this out](#)

MLE solutions

Bernoulli

► $p = \frac{n_1}{n}$ where $n_1 = \sum_{i=1}^n x_i$

Categorical

► $\theta_x = \frac{n_x}{n}$ where $n_x = \sum_{i=1}^n \delta_{x_i x}$
for all $x \in \mathcal{X} = \{1, \dots, k\}$

► Quiz

δ is the Kronecker delta



MLE: Bernoulli

Probability mass function

$$\blacktriangleright f(a; p) = \text{Bern}(X = a|p) = p^a(1 - p)^{1-a}$$

Problem: optimisation of the log-likelihood function $\mathcal{L}(p)$

$$p^* = \underset{p \in (0,1)}{\operatorname{argmax}} \underbrace{\sum_{i=1}^n \log f(x_i; p)}_{\mathcal{L}(p)}$$

Strategy

1. set first derivative of $\mathcal{L}(p)$ to 0
2. solve for p

Bernoulli: MLE derivation

Derivative

$$\begin{aligned}\frac{d\mathcal{L}(p)}{dp} &= \frac{d}{dp} \left[\sum_{i=1}^n x_i \log p + (1 - x_i) \log(1 - p) \right] \\&= \sum_{i=1}^n x_i \frac{d}{dp} \log p + (1 - x_i) \frac{d}{dp} \log(1 - p) \\&= \sum_{i=1}^n \frac{x_i}{p} + \frac{1 - x_i}{1 - p} (-1) \\&= \sum_{i=1}^n \frac{x_i(1 - p) - (1 - x_i)p}{p(1 - p)} \\&= \frac{(1 - p)}{p(1 - p)} \underbrace{\sum_{i=1}^n x_i}_{n_1} - \frac{p}{p(1 - p)} \underbrace{\sum_{i=1}^n 1 - x_i}_{n_0} \\&= \frac{(1 - p)}{p(1 - p)} n_1 - \frac{p}{p(1 - p)} n_0\end{aligned}$$

Set to 0 and solve for p

$$\begin{aligned}0 &= \frac{(1 - p)}{p(1 - p)} n_1 - \frac{p}{p(1 - p)} n_0 \\&= (1 - p) n_1 - p n_0 \\&= n_1 - p n_1 - p n_0 \\&= n_1 - p(n_1 + n_0) \\n_1 &= p(n_1 + n_0) \\p &= \frac{n_1}{n_1 + n_0} \\p &= \frac{n_1}{n}\end{aligned}$$

Note

- ▶ $n_1 = \sum_{i=1}^n x_i$
- ▶ $n_0 = \sum_{i=1}^n (1 - x_i)$
- ▶ $n = n_1 + n_0$

MLE: Categorical

Probability mass function

► Categorical

$$f(a; \theta_1^k) = \text{Cat}(X = a | \theta_1, \dots, \theta_k) = \prod_{x=1}^k \theta_x^{\delta_{xa}}$$

Problem: optimisation of the log-likelihood function $\mathcal{L}(\theta_1^k)$

$$p^* = \underset{\theta_1^k \in \mathbb{R}_{>0}}{\text{argmax}} \underbrace{\sum_{i=1}^n \log f(x_i; \theta_1^k)}_{\mathcal{L}(\theta_1, \dots, \theta_k)} \quad \text{s.t.} \quad \sum_{x=1}^k \theta_x = 1$$

Strategy

1. introduce Lagrange multiplier λ for the constraint $\sum_{x=1}^k \theta_x = 1$
2. set partial derivatives to 0
3. solve for λ and θ_1^k

Check the complete derivation



Next week

Lab2

- ▶ probability theory
- ▶ MLE for Bernoulli and Categorical

Next lecture we will discuss sequence prediction

- ▶ we will model with Categorical distributions
- ▶ and obtain maximum likelihood estimates from text

References I