Modeling data with probability distributions

DSE 210

Distributional modeling

A useful way to summarize a data set:

- Fit a probability distribution to it.
- Simple and compact, and captures the big picture while smoothing out the wrinkles in the data.
- In subsequent application, use distribution as a proxy for the data.

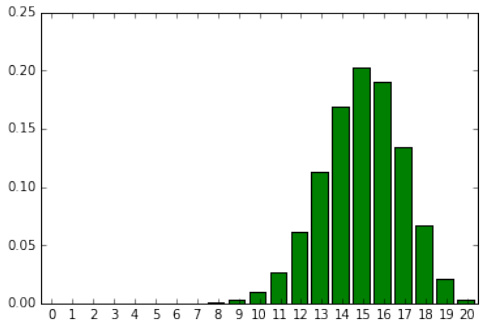
Which distributions to use?

> *There exist a few distributions of great universality which occur in a surprisingly large number of problems. The three principal distributions, with ramifications throughout probability theory, are the binomial distribution, the normal distribution, and the Poisson distribution.* – William Feller.

Well, this is true in one dimension. For higher-dimensional data, we'll use combinations of 1-d models: **products** and **mixtures**.

The binomial distribution

Binomial(n, p): the number of heads when n coins of bias (heads probability p) are tossed, independently.



Suppose X has a binomial(n, p) distribution.

$$\mathbb{E}X = np$$
$$\text{var}(X) = np(1-p)$$
$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Fitting a binomial distribution to data

Example: Upcoming election in a two-party country.

- You choose 1000 people at random and poll them.
- 600 say Democratic.

What is a good estimate for the fraction of votes the Democrats will get in the election? Clearly, 60%.

More generally, you observe n tosses of a coin of unknown bias. k of them are heads. How to estimate the bias?

$$p = \frac{k}{n}$$

Maximum likelihood estimation

Let $\mathcal{P}$ be a class of probability distributions (Gaussians, Poissons, etc).

Maximum likelihood principle: pick the distribution in $\mathcal{P}$ that makes the data maximally likely.
That is, pick the $p \in \mathcal{P}$ that maximizes $\Pr(\text{data}|p)$.

E.g. Suppose $\mathcal{P}$ is the class of binomials. We observe n coin tosses, and k of them are heads.

- Maximum likelihood : pick the bias p that maximizes

$$\Pr(\text{data}|p) = p^k(1-p)^{n-k}.$$

- Maximizing this is the same as maximizing its log,

$$\text{LL}(p) = k \ln p + (n-k)\ln(1-p).$$

- Set the derivative to zero.

$$\text{LL}'(p) = \frac{k}{p} - \frac{n-k}{1-p} = 0 \quad \Rightarrow \quad p = \frac{k}{n}.$$

Maximum likelihood: a small caveat

You have two coins of unknown bias.

- You toss the first coin 10 times, and it comes out heads every time. You estimate its bias as $p_1 = 1.0$.
- You toss the second coin 10 times, and it comes out heads once. You estimate its bias as $p_2 = 0.1$.

Now you are told that one of the coins was tossed 20 times and 19 of them came out heads. Which coin do you think it is?

- Likelihood under p_1:

$$\Pr(\text{19 heads out of 20 tosses}|\text{bias} = 1) = 0$$

- Likelihood under p_2:

$$\Pr(\text{19 heads out of 20 tosses}|\text{bias} = 0.1) = (0.1)^{19}(0.9)^1$$

The likelihood principle would choose the second coin. Is this right?

Laplace smoothing

A smoothed version of maximum-likelihood: when you toss a coin n times and observe k heads, estimate the bias as

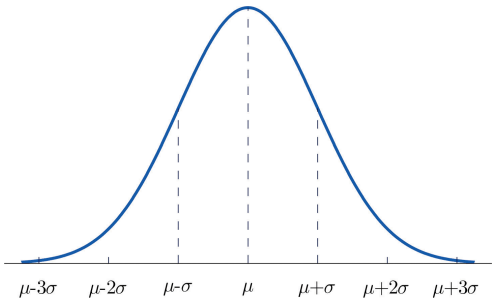
$$p = \frac{k+1}{n+2}.$$

Laplace's law of succession: What is the probability that the sun won't rise tomorrow?

- Let p be the probability that the sun won't rise on a randomly chosen day. We want to estimate p.
- For the past 5000 years (= 1825000 days), the sun has risen every day. Using Laplace smoothing, estimate

$$p = \frac{1}{1825002}.$$

The normal distribution



The normal (or *Gaussian*) $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- 66% of the distribution lies within one standard deviation of the mean, i.e. in the range $\mu \pm \sigma$
- 95% lies within $\mu \pm 2\sigma$
- 99% lies within $\mu \pm 3\sigma$

Maximum likelihood estimation of the normal

Suppose you see n data points $x_1, \ldots, x_n \in \mathbb{R}$, and you want to fit a Gaussian $N(\mu, \sigma^2)$ to them. How to choose μ, σ?

- Maximum likelihood: pick μ, σ to maximize

$$\Pr(\text{data}|\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right)$$

- Work with the log, since it makes things easier:

$$\text{LL}(\mu, \sigma^2) = \frac{n}{2} \ln \frac{1}{2\pi\sigma^2} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

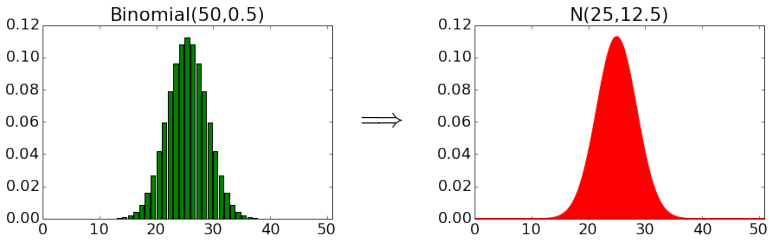
- Setting the derivatives to zero, we get

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

These are simply the empirical mean and variance.

Normal approximation to the binomial



When a coin of bias p is tossed n times, let X be the number of heads.

- We know X has mean np and variance $np(1-p)$.
- As n grows, the distribution of X looks increasingly like a Gaussian with this mean and variance.

Application to sampling

We want to find out what fraction p of San Diegans know how to surf. So we poll n random people, and find that k of them surf. Our estimate:

$$\widehat{p} = \frac{k}{n}.$$

Normal approximation:

- k has a binomial(n, p) distribution.
- This is close to a Gaussian with mean np and variance $np(1-p)$.
- Therefore the distribution of $\widehat{p} = k/n$ is close to a Gaussian with

$$\text{mean} = p$$

$$\text{variance} = \frac{p(1-p)}{n} \leq \frac{1}{4n}$$

Confidence intervals:

- With 95% confidence, our estimate is accurate within $\pm 1/\sqrt{n}$.
- With 99% confidence, our estimate is accurate within $\pm 3/2\sqrt{n}$.

The multinomial distribution

A k-sided die:

- A fair coin has two possible outcomes, each equally likely.
- A fair die has six possible outcomes, each equally likely.
- Imagine a k-faced die, with probabilities $p_1, \ldots, p_k$.

Toss such a die n times, and count the number of times each of the k faces occurs:

$$X_j = \#\text{ of times face } j \text{ occurs}$$

The distribution of $X = (X_1, \ldots, X_k)$ is called the **multinomial**.

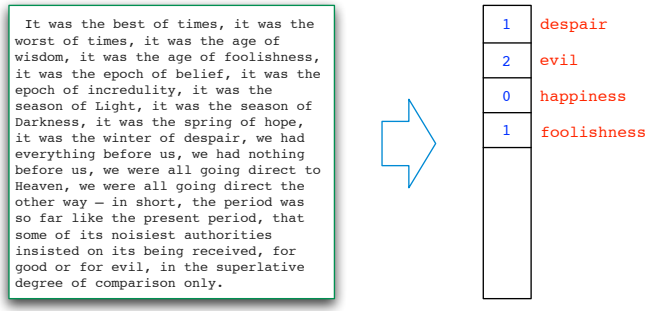
- Parameters: $p_1, \ldots, p_k \geq 0$, with $p_1 + \cdots + p_k = 1$.
- $\mathbb{E}X = (np_1, np_2, \ldots, np_k)$.
- $\Pr(n_1, \ldots, n_k) = \binom{n}{n_1, n_2, \ldots, n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$, where

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$$

is the number of ways to place balls numbered $\{1, \ldots, n\}$ into bins numbered $\{1, \ldots, k\}$.

Example: text documents

Bag-of-words: vectorial representation of text documents.

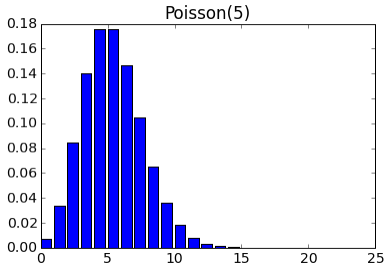


- Fix V = some vocabulary.
- Treat the words in a document as independent draws from a multinomial distribution over V:

$$p = (p_1, \ldots, p_{|V|}), \quad \text{such that} \quad p_i \geq 0 \text{ and } \sum_i p_i = 1$$

The Poisson distribution

A distribution over the non-negative integers $\{0, 1, 2, \ldots\}$



The Poisson has parameter $\lambda > 0$, with $\Pr(X = k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$

- Mean: $\mathbb{E}X = \lambda$
- Variance: $\mathbb{E}(X - \lambda)^2 = \lambda$
- Maximum likelihood fit: set λ to the empirical mean

How the Poisson arises

Count the number of events (collisions, phone calls, etc) that occur in a certain interval of time. Call this number X, and say it has expected value λ.

Now suppose we divide the interval into small pieces of equal length.

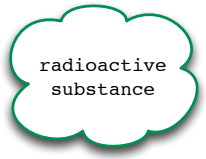
If the probability of an event occurring in a small interval is:

- independent of what happens in other small intervals, and
- the same across small intervals,

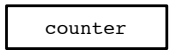
then $X \sim \text{Poisson}(\lambda)$.

Poisson: examples

Rutherford's experiments with radioactive disintegration (1920)

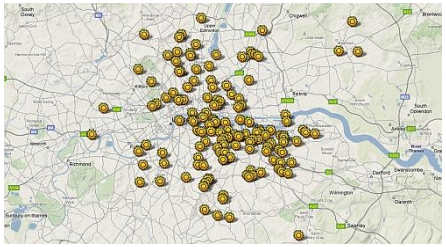


- $N = 2608$ intervals of 7.5 seconds
- N_k = # intervals with k particles
- Mean: 3.87 particles per interval



k	0	1	2	3	4	5	6	7	8	≥ 9
N_k	57	203	383	525	532	408	273	139	45	43
P(3.87)	54.4	211	407	526	508	394	254	140	67.9	46.3

Flying bomb hits on London in WWII



- Area divided into 576 regions, each 0.25 km^2
- N_k = # regions with k hits
- Mean: 0.93 hits per region

k	0	1	2	3	4	≥ 5
N_k	229	211	93	35	7	1
P(0.93)	226.8	211.4	98.54	30.62	7.14	1.57

Multivariate distributions

Almost all distributions we've considered are for one-dimensional data.

- Binomial, Poisson: integer
- Gaussian: real

What to do with the usual situation of data in higher dimensions?

1. Model each coordinate separately and treat them as independent.
   For $x = (x_1, \ldots, x_p)$, fit separate models $\Pr_i$ to each x_i, and assume

   $$\Pr(x_1, \ldots, x_p) = \Pr_1(x_1)\Pr_2(x_2)\cdots\Pr_p(x_p).$$

   This assumption is almost always completely inaccurate, and sometimes causes problems.
2. Multivariate Gaussian.
   Allows modeling of correlations between coordinates.
3. More general graphical models.
   Arbitrary dependencies between coordinates.