Lecture 2: Statistical inferences

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center
Office hours: by appointment
Statistical inferences in NLP
Authorship attribution

Given two data sets $D_1$ and $D_2$
(e.g. the known works of Shakespeare and of Marlowe)
where does the new data set $D'$ come from?
(e.g. a disputed piece)

Assume $D_1 \sim \theta_1$ and $D_2 \sim \theta_2$
Each set is generated by a different underlying distribution

If $P(D' \mid \theta_1) > P(D' \mid \theta_2)$, assume $D'$ is more like $D_1$

This requires us to estimate the parameters $\theta$
from the data $D$
Computing $P(D \mid \theta)$

We are given a data set $D$ with $n$ items $D = (x_1, \ldots, x_n)$

We assume $D$ is generated from a distribution with parameters $\theta$

What is the probability of $D$?

We assume the items $x_i$ are independent and identically distributed (i.i.d.):

$$x_i \sim P(D \mid \theta) = P(x_1, \ldots, x_n \mid \theta) = \prod_{i=1..n} P(x_i \mid \theta)$$

$= \text{We assume the } x_i \text{ are exchangeable}$
Statistical inferences (I)

We are given a data set \( D \) with \( n \) items \( D = (x_1, \ldots, x_n) \).

We assume \( D \) is generated (sampled) from an (unknown) distribution with parameters \( \theta \): \( x_i \sim \theta \)
\( \theta \): the parameters of a probability distribution

What is the **probability of the next item**?
\[
x_{n+1} = \sum_i P(x \mid x_1 \ldots x_n)
\]

What is the **most likely next item**?
\[
x^*_{n+1} = \arg\max_x P(x \mid x_1 \ldots x_n)
\]

This requires the **predictive distribution** \( P(x_{n+1} \mid x_1 \ldots x_n) \)

NLP applications: language modeling
We may also be given a data set $D$ with $n$ items
$$D = ((x_1, y_1), \ldots, (x_n, y_n))$$
and need to know the **most likely hidden value** $y_{n+1}$
for a previously unseen item $x_{n+1}$
$$y_{n+1} = \text{argmax}_y P(y | x_{n+1}; D)$$
(Supervised learning)

NLP applications:
- POS-tagging, Parsing, sentiment analysis, etc.
Statistical inferences (III)

Or, we may be given in **incomplete** data set

\[ D = ((x_1, _), ...., (x_n, _)) \]

and need to know the **most likely hidden value** \( y_{n+1} \) for a previously unseen item \( x_{n+1} \)

\[ y_{n+1} = \text{argmax}_y P(y | x_{n+1}; D) \]

(= Unsupervised learning)

**Common notation:** \( x_i \) is observed, \( y_i \) is hidden
Or, we may be given in **incomplete** data set

\[ D = ((x_1, _), \ldots, (x_n, _)) \]

where there are **latent** variables \( z_i \):

\[ (x_i, z_i) \sim \theta \]

We need to assign probabilities to \( x_{n+1} \),
or find the **most likely** \( x_{n+1} \)

\[ P(x \mid x_{n+1}; D) \]

\[ x_{n+1} = \text{argmax}_x P(x \mid x_{n+1}; D) \]

= (one kind of) partially supervised learning
Or, we may be given in incomplete data set $D = ((x_1, y_1), \ldots, (x_n, y_n))$

where there are latent variables $z_i$: $(x_i, y_i, z_i) \sim \theta$

We need to know the most likely $y_{n+1}$ for $x_{n+1}$

$y_{n+1} = \arg\max_y P(y | x_{n+1}; D)$

= (one kind of) partially supervised learning
Bayesian statistics
Bayesian statistics

\( \theta \): the parameters of a probability distribution

Probabilities represent degrees of belief

Data \( D \) provide evidence for/against our beliefs.

We update our belief \( \theta \) based on evidence we see:

\[
P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta) \, d\theta}
\]

For fixed data \( D \), \( P(D|\theta) \) is the **likelihood** of \( \theta \)
Bayesian statistics

\[
P(\theta | D) = \frac{P(\theta) P(D | \theta)}{\int P(\theta) P(D | \theta) d\theta}
\]

- \(P(\theta | D)\): Posterior Probability of \(\theta\)
- \(P(\theta)\): Prior Probability of \(\theta\)
- \(P(D | \theta)\): Likelihood of \(D\)
- \(P(D) = \text{Marginal Likelihood of } D\)
Bayesian statistics

The posterior $P(\theta \mid D)$ is proportional to the prior $P(\theta)$ times the likelihood $P(D \mid \theta)$:

$$P(\theta \mid D) \propto P(\theta)P(D \mid \theta)$$
Discrete probability distributions:
Throwing a coin

Bernoulli distribution:
- Probability of success (=head,yes) in single yes/no trial
  - The probability of head is $p$.
  - The probability of tail is $1-p$.

Binomial distribution:
- Prob. of the number of heads in a sequence of yes/no trials
- The probability of getting exactly $k$ heads in $n$ independent yes/no trials is:
  $$ P(k \text{ heads, } n-k \text{ tails}) = \binom{n}{k} p^k (1-p)^{n-k} $$
Looking at the binomial distribution again
The binomial distribution

If $p$ is the probability of heads, the probability of getting exactly $k$ heads in $n$ independent yes/no trials is given by the binomial distribution $Bin(n,p)$:

$$P(k \text{ heads}) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$= \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k}$$

Expectation $E(Bin(n,p)) = np$

Variance $var(Bin(n,p)) = np(1-p)$
Parameter estimation

Given data $D=HTTHTT$, what is the probability $\theta$ of heads?

**Maximum likelihood estimation (MLE):**
Use the $\theta$ which has the highest likelihood $P(D|\theta)$.

$$\theta_{MLE} = \arg \max_{\theta} P(D|\theta)$$

**Maximum a posterior estimation (MAP):**
Use the $\theta$ which has the highest posterior probability $P(\theta|D)$.

$$\theta_{MAP} = \arg \max_{\theta} P(\theta|D) = \arg \max_{\theta} P(\theta)P(D|\theta)$$

**Bayesian estimation:**
Integrate over all $\theta$ $\Rightarrow$ Compute the expectation of $\theta$ given $D$:

$$P(x = H|D) = \int_{0}^{1} P(x = H|\theta)P(\theta|D)d\theta = E[\theta|D]$$
Binomial likelihood

What distribution does $p$ (probability of heads) have, given that the data $D$ consists of $\#H$ heads and $\#T$ tails?

Likelihood $L(\theta; D=(\#\text{Heads}, \#\text{Tails}))$ for binomial distribution
Maximum likelihood estimation for the coin flip

\[
\theta^* = \arg\max_\theta P(D|\theta) \\
= \arg\max_\theta \theta^H (1 - \theta)^T \\
= \frac{H}{H + T}
\]
Bayesian estimation: what prior?

The posterior $P(\theta | D)$ is proportional to prior x likelihood:

$$P(\theta | D) \propto P(\theta) P(D | \theta)$$

The likelihood $P(D | \theta)$ of a binomial is $P(D | \theta) = \theta^H (1-\theta)^T$

Assume the prior $P(\theta)$ is proportional to powers of $\theta$ and $(1-\theta)$:

$$P(\theta) \propto \theta^a (1-\theta)^b$$

Then the posterior $P(\theta | D)$ will also be proportional to powers of $\theta$ and $(1-\theta)$:

$$P(\theta | D) \propto P(\theta) P(D | \theta)$$

$$= \theta^a (1-\theta)^b \theta^H (1-\theta)^T$$

$$= \theta^{a+H} (1-\theta)^{b+T}$$
In search of a prior for coin flips...

We would like something of the form:

\[ P(\theta) \propto \theta^a (1 - \theta)^b \]

But -- this looks just like the binomial:

\[ P(k \text{ heads}) = \binom{n}{k} p^k (1 - p)^{n-k} \]

\[ = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k} \]

…. except that \( k \) is an integer and \( \theta \) is a real with \( 0 < \theta < 1 \).
The Gamma function

The Gamma function $\Gamma(x)$ is the generalization of the factorial $x!$ (or rather $(x-1)!$) to the reals:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx \quad \text{for } \alpha > 0$$

For $x > 1$, $\Gamma(x) = (x-1)\Gamma(x-1)$.

For positive integers, $\Gamma(x) = (x-1)!$
The Gamma function
The Beta distribution

A random variable $X \ (0 < x < 1)$ has a Beta distribution with (hyper)parameters $\alpha \ (\alpha > 0)$ and $\beta \ (\beta > 0)$ if $X$ has a continuous distribution with probability density function

$$P(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

The first term is a normalization factor (to obtain a distribution)

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \, dx = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Expectation: $\frac{\alpha}{\alpha + \beta}$
Beta as prior for binomial

Given a prior \( P(\theta | \alpha, \beta) = \text{Beta}(\alpha, \beta) \), and data \( D=(H,T) \), what is our posterior?

\[
P(\theta | \alpha, \beta, H, T) \propto P(H, T | \theta) P(\theta | \alpha, \beta)
\]

\[
\propto \theta^H (1 - \theta)^T \theta^{\alpha-1} (1 - \theta)^{\beta-1}
\]

\[
= \theta^{H+\alpha-1} (1 - \theta)^{T+\beta-1}
\]

With normalization

\[
P(\theta | \alpha, \beta, H, T) = \frac{\Gamma(H + \alpha + T + \beta)}{\Gamma(H + \alpha) \Gamma(T + \beta)} \theta^{H+\alpha-1} (1 - \theta)^{T+\beta-1}
\]

\[
= \text{Beta}(\alpha + H, \beta + T)
\]
So, what do we predict?

Our Bayesian estimate for the next coin flip $P(x=1 \mid D)$:

$$P(x = H \mid D) = \int_0^1 P(x = H \mid \theta) P(\theta \mid D) d\theta$$
**Beta(α, β) with α > 1, β > 1:** unimodal
\( \text{Beta}(\alpha, \beta) \) with \( \alpha < 1, \beta < 1 \): U-shaped

Beta Distribution \( \text{Beta}(\alpha, \beta) \)
\textit{Beta}(\alpha, \beta) \text{ with } \alpha = \beta: \text{ symmetric} \\
(\alpha = \beta = 1: \text{ uniform})
**Beta(α,β)** with α<1, β >1: strictly decreasing
**Beta(\(\alpha, \beta\)) with \(\alpha = 1, \beta > 1\)**

- \(\alpha = 1, 1 < \beta < 2\): strictly concave.
- \(\alpha = 1, \beta = 2\): straight line
- \(\alpha = 1, \beta > 2\): strictly convex
Conjugate priors

The beta distribution is a \textit{conjugate prior} to the binomial: the resulting posterior is also a beta distribution.

All members of the \textit{exponential family} of distributions have conjugate priors.

Examples:
- Multinomial: conjugate prior = Dirichlet
- Gaussian: conjugate prior = Gaussian
Conjugate priors

The **posterior** is proportional to **prior x likelihood**:

\[ P(\theta | D) \propto P(\theta) \ P(D|\theta) \]

**Conjugate priors:**
Posterior is the same kind of distribution as prior.

For **binomial likelihood**:
conjugate prior = Beta distribution
Discrete probability distributions: Rolling a die

Categorical distribution:
Probability of getting one of $N$ outcomes in a single trial.
The probability of category/outcome $c_i$ is $p_i$ ($\sum p_i = 1$)

Multinomial distribution:
Probability of observing each possible outcome $c_i$ exactly $X_i$ times in a sequence of $n$ trials

\[
P(X_1 = x_i, \ldots, X_N = x_N) = \frac{n!}{x_1! \cdots x_N!} p_1^{x_1} \cdots p_N^{x_N} \quad \text{if} \quad \sum_{i=1}^{N} x_i = n
\]
Moving on to multinomials
Multinomials have a Dirichlet prior

Multinomial distribution:
Probability of observing each possible outcome \( c_i \) exactly \( X_i \) times in a sequence of \( n \) trials:

\[
P(X_1 = x_i, \ldots, X_K = x_k) = \frac{n!}{x_1! \cdots x_K!} \theta_1^{x_1} \cdots \theta_K^{x_K} \quad \text{if} \quad \sum_{i=1}^{N} x_i = n
\]

Dirichlet prior:

\[
Dir(\theta | \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_K)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}
\]
Multinomial variables

- In NLP, $X$ is often a **discrete** random variable that can take one of $K$ states.

- We can represent such $X$s as *$K$-dimensional vectors* in which one $x_k = 1$ and all other elements are 0
  
  $x = (0,0,1,0,0)^T$

- Denote probability of $x_k = 1$ as $\mu_k$ with $0 \leq \mu_k \leq 1$ and $\sum_k \mu_k = 1$

Then the probability of $x$ is:

$$P(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$
Multinomial likelihood

What is the **likelihood** of \( D = x_1 \ldots x_i \ldots x_N \)?

Define

\[
    m_k := \sum_{n=1}^{N} x_{nk}
\]

(= #observations with \( x_k = 1 \))

The likelihood depends only on the \( m_k \).

\( m_k \) are **sufficient statistics**
Multinomials: Dirichlet prior

The joint distribution of \((m_1, \ldots, m_K)\) conditioned on \(\mu\) and \(N\) is a **multinomial distribution**:

\[
P(m_1, \ldots, m_K = x_k) = \frac{N!}{m_1! \cdots m_K!} \theta_1^{m_1} \cdots \theta_K^{x_K}
\]

if \(\sum_{i=1}^{K} x_k = N\)

**Multinomials** have a **Dirichlet prior** with hyperparameters \(\alpha\):

\[
Dir(\theta|\alpha_1, \ldots \alpha_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k-1}
\]
The Dirichlet

A Dirichlet is confined to a simplex (here $\mu=(\mu_1, \mu_2, \mu_3)$)

(Figure from Chris Bishop’s PRML book & website)
Examples of the Dirichlet

\[ \{\alpha_k\} = 0.1 \quad \{\alpha_k\} = 1 \quad \{\alpha_k\} = 10 \]

(all figures from Chris Bishop’s PRML book & website)
Dirichlet as conjugate prior

Given a prior $\text{Dir}(\mu | \alpha)$ and Data $D$ with sufficient statistics $m = (m_1, ..., m_K)$, the posterior is

$$p(\mu | D, \alpha) \propto P(D | \mu) P(\mu)$$

$$\propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1 + m_k}$$

The normalized posterior is:

$$p(\mu | D, \alpha) = \text{Dir}(\mu | \alpha + m)$$

$$= \frac{\Gamma(\alpha_1 + \ldots + \alpha_K + N)}{\Gamma(\alpha_1 + m_1) \times \ldots \times \Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1 + m_k}$$
Likelihood, prior and posterior for the Dirichlet/multinomial

Likelihood $P(Y|\theta) = \prod_{k=1}^{K} \theta_{k}^{m_k}$

Prior $P(\theta|\alpha) \propto \prod_{k=1}^{K} \theta_{k}^{\alpha_k - 1}$

Posterior $P(\theta|Y, \alpha) \propto \prod_{k=1}^{K} \theta_{k}^{m_k + \alpha_k - 1}$
MLE vs Bayesian estimate

**Maximum likelihood estimate:**
Maximize $\ln p(D|\mu)$ wrt. $\mu_k$ under the constraint that $\sum \mu_k = 1$

(...Use Lagrange multipliers...)

$$\mu_k^\text{MLE} = \frac{m_k}{N}$$

**Bayesian estimate:**

$$\mu_k^\text{BE} = \frac{m_k + \alpha_k}{N + \sum_{k'=1}^{K} \alpha_{k'}}$$
More about conjugate priors

- We can interpret the hyperparameters as “pseudocounts”

- Sequential estimation (updating counts after each observation) gives same results as batch estimation

- Add-one smoothing (Laplace smoothing) = uniform prior

- On average, more data leads to a sharper posterior (sharper = lower variance)
Today’s reading

- Bishop, Pattern Recognition and Machine Learning, Ch. 2