Applied Machine Learning

Maximum Likelihood and Bayesian Reasoning

Siamak Ravanbakhsh

COMP 551 **(fall 2020)**

Objectives

understand what it means to learn a probabilistic model of the data

- using maximum likelihood principle
- using Bayesian inference
  - prior, posterior, posterior predictive
  - MAP inference
  - Beta-Bernoulli conjugate pairs

Parameter estimation

a thumbtack's head/tail outcome has a **Bernoulli distribtion**

$$\text{Bernoulli}(x|\theta) = \theta^x(1-\theta)^{(1-x)}$$

this is our **probabilistic model** of some head/tail IID data $\mathcal{D} = \{0, 0, 1, 1, 0, 0, 1, 0, 0, 1\}$

Objective: learn the model parameter θ

since we are only interested in the counts, we can also use **Binomial distribution**

$$\text{Binomial}(N, N_h|\theta) = \binom{N}{N_h}\theta^{N_h}(1-\theta)^{N-N_h}$$

$|\mathcal{D}|$ — # heads $N_h = \sum_{x\in\mathcal{D}} x$ — N_t

Maximum likelihood

a thumbtack's head/tail outcome has a **Bernoulli distribtion**

$$\text{Bernoulli}(x|\theta) = \theta^x(1-\theta)^{(1-x)}$$

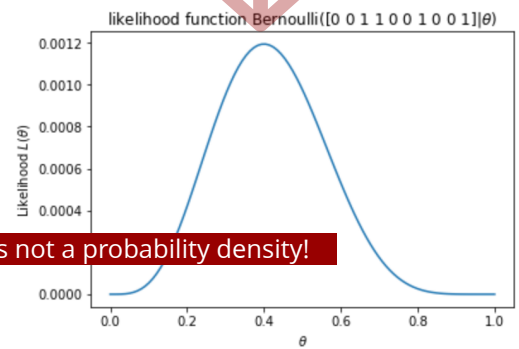
this is our **probabilistic model** of some head/tail IID data $\mathcal{D} = \{0, 0, 1, 1, 0, 0, 1, 0, 0, 1\}$

Objective: **learn** the model parameter θ

Idea: find the parameter θ that maximizes the probability of observing $\mathcal{D}$

Likelihood $L(\theta; \mathcal{D}) = \prod_{x \in \mathcal{D}} \text{Bernoulli}(x|\theta) = \theta^4(1-\theta)^6$ is a function of θ

Max-likelihood assignment



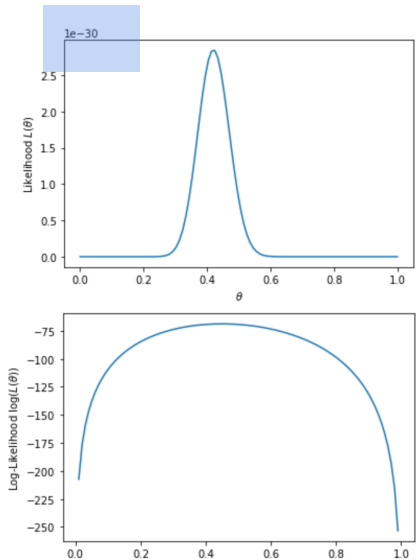
note that this is not a probability density!

Maximizing log-likelihood

likelihood $L(\theta; \mathcal{D}) = \prod_{x \in \mathcal{D}} p(x; \theta)$

using product here creates extreme values

for 100 samples in our example, the likelihood shrinks below 1e-30



log-likelihood has the same maximum but it is well-behaved

$$\ell(\theta; \mathcal{D}) = \log(L(\theta; \mathcal{D})) = \sum_{x \in \mathcal{D}} \log(p(x; \theta))$$

how do we find the max-likelihood parameter? $\theta^* = \arg\max_\theta \ell(\theta; \mathcal{D})$

*for some simple models we can get the **closed form solution***

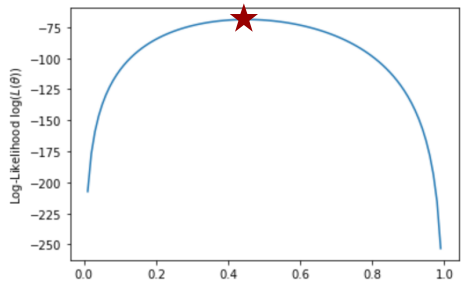
*for complex models we need to use **numerical optimization***

Maximizing log-likelihood

log-likelihood $\ell(\theta; \mathcal{D}) = \log(L(\theta; \mathcal{D})) = \sum_{x \in \mathcal{D}} \log(\text{Bernoulli}(x; \theta))$

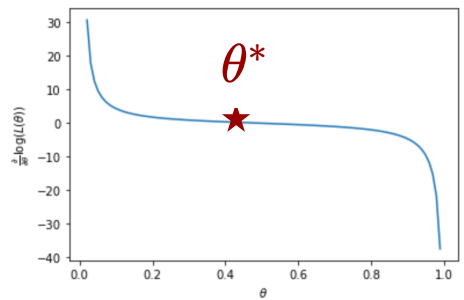
observation: at maximum, the derivative of $\ell(\theta; \mathcal{D})$ is zero

idea: set the the derivative to zero and solve for θ



example max-likelihood for Bernoulli

$$\begin{aligned}\frac{\partial}{\partial\theta}\ell(\theta;\mathcal{D}) &= \frac{\partial}{\partial\theta}\sum_{x\in\mathcal{D}}\log\left(\theta^x(1-\theta)^{(1-x)}\right)\\ &= \frac{\partial}{\partial\theta}\sum_x x\log\theta + (1-x)\log(1-\theta)\\ &= \sum_x \frac{x}{\theta} - \frac{1-x}{1-\theta} = 0\end{aligned}$$



which gives $\theta^{MLE} = \frac{\sum_{x\in\mathcal{D}} x}{|\mathcal{D}|}$ is simply the portion of heads in our dataset

Bayesian approach

max-likelihood estimate does not reflect our uncertainty:

- e.g., $\theta^* = .2$ for both 1/5 heads and 1000/5000 heads

in the Bayesian approach

- we maintain a *distribution* over parameters $p(\theta)$ prior
- after observing $\mathcal{D}$ we update this distribution $p(\theta|\mathcal{D})$ posterior

how to do this update? using **Bayes rule**

$$p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})}$$

prior

likelihood of the data
previously denoted by $L(\theta; \mathcal{D})$

evidence: this is a normalization $p(\mathcal{D}) = \int p(\theta)p(\mathcal{D}|\theta)\mathrm{d}\theta$

Conjugate Priors

in our running example, we know the form of likelihood:

prior	$p(\theta)?$
posterior	$p(\theta \mid \mathcal{D})?$
likelihood	$p(\mathcal{D} \mid \theta) = \prod_{x \in \mathcal{D}} \text{Bernoulli}(x; \theta) \quad = \theta^{N_h}(1-\theta)^{N_t}$

we want prior and posterior to have the **same form** (so that we can easily update our belief with new observations.)

this gives us the following form $p(\theta \mid a, b) \propto \theta^a (1-\theta)^b$

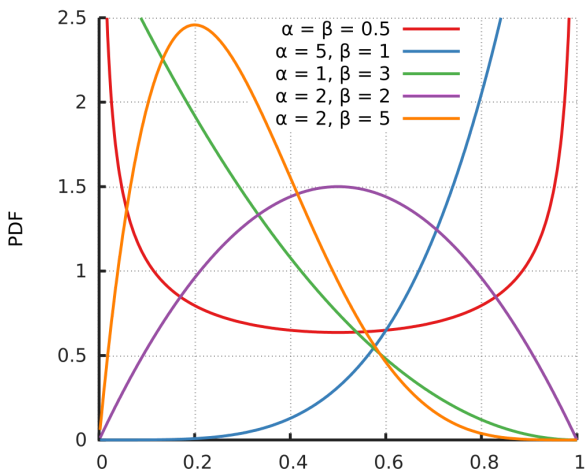
↓ this means there is a normalization constant that does not depend on θ

distribution of this form has a name, **Beta** distribution

we say Beta distribution is a conjugate prior to the Bernoulli likelihood

Beta distribution

Beta distribution has the following density



$$\text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Γ normalization
is the generalization of factorial to real number $\Gamma(a+1) = a\Gamma(a)$

$\alpha, \beta > 0$

$\text{Beta}(\theta|\alpha=\beta=1)$ is uniform

mean of the distribution is $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$

for $\alpha, \beta > 1$ the dist. is unimodal; its mode is $\frac{\alpha-1}{\alpha+\beta-2}$

Beta-Bernoulli conjugate pair

prior $p(\theta) = \text{Beta}(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

likelihood $p(\mathcal{D}|\theta) = \theta^{N_h}(1-\theta)^{N_t}$ | *product of Bernoulli likelihoods*
equivalent to Binomial likelihood

posterior $p(\theta|\mathcal{D}) = \text{Beta}(\theta|\alpha + N_h, \beta + N_t) \propto \theta^{\alpha+N_h-1}(1-\theta)^{\beta+N_t-1}$

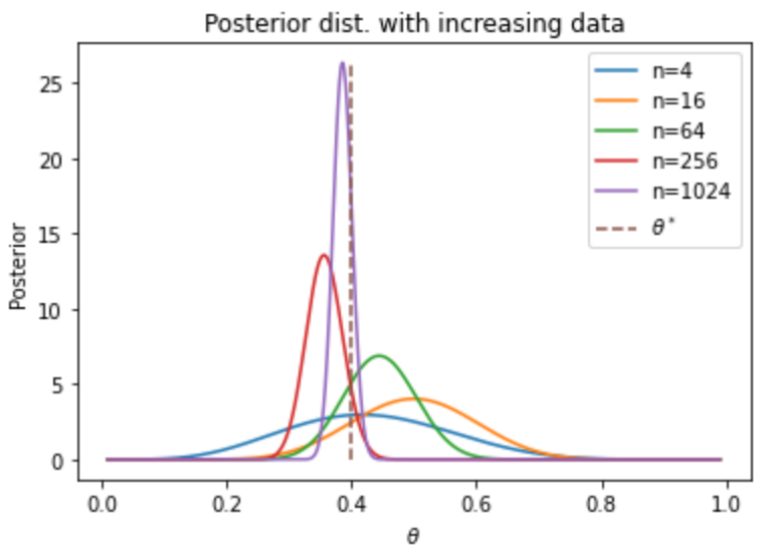
α, β are called *pseudo-counts*
their effect is similar to imaginary observation of heads (α) and tails (β)

Effect of more data

with few observations, prior has a high influence

as we increase the number of observations $N = |\mathcal{D}|$ the effect of prior diminishes

the likelihood term dominates the posterior



example prior $p(\theta; 5, 5)$

plot of the posterior density with **n** observations

$$p(\theta|\mathcal{D}) \propto \theta^{5+H}(1-\theta)^{5+N-H}$$

Posterior predictive

our goal was to estimate the parameters (θ) so that we can make predictions $p(x|\theta)$

but now we have a (posterior) **distribution** over parameters $p(\theta|\mathcal{D})$

rather than using a single parameter $p(x|\theta)$

we need to calculate the average prediction

$$p(x|\mathcal{D}) = \int_\theta p(\theta|\mathcal{D})p(x|\theta)\mathrm{d}\theta$$ posterior predictive

for each possible θ, weight the prediction by the posterior probability of that parameter being true

Posterior predictive for Beta-Bernoulli

start from a Beta prior $p(\theta) = \mathrm{Beta}(\theta|\alpha, \beta)$

observe N_h heads and N_t tails, the posterior is $p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|\alpha + N_h, \beta + N_t)$

what is the probability that the next coin flip is head?

$$p(x = 1|\mathcal{D}) = \int_\theta \mathrm{Bernoulli}(x = 1|\theta)\mathrm{Beta}(\theta|\alpha + N_h, \beta + N_t)\mathrm{d}\theta$$

$$= \underbrace{\int_\theta \theta\, \mathrm{Beta}(\theta|\alpha + N_h, \beta + N_t)}_{\text{mean of Beta dist.}} = \frac{\alpha + N_h}{\alpha + \beta + N}$$

compare with prediction of maximum-likelihood: $p(x = 1|\mathcal{D}) = \frac{N_h}{N}$

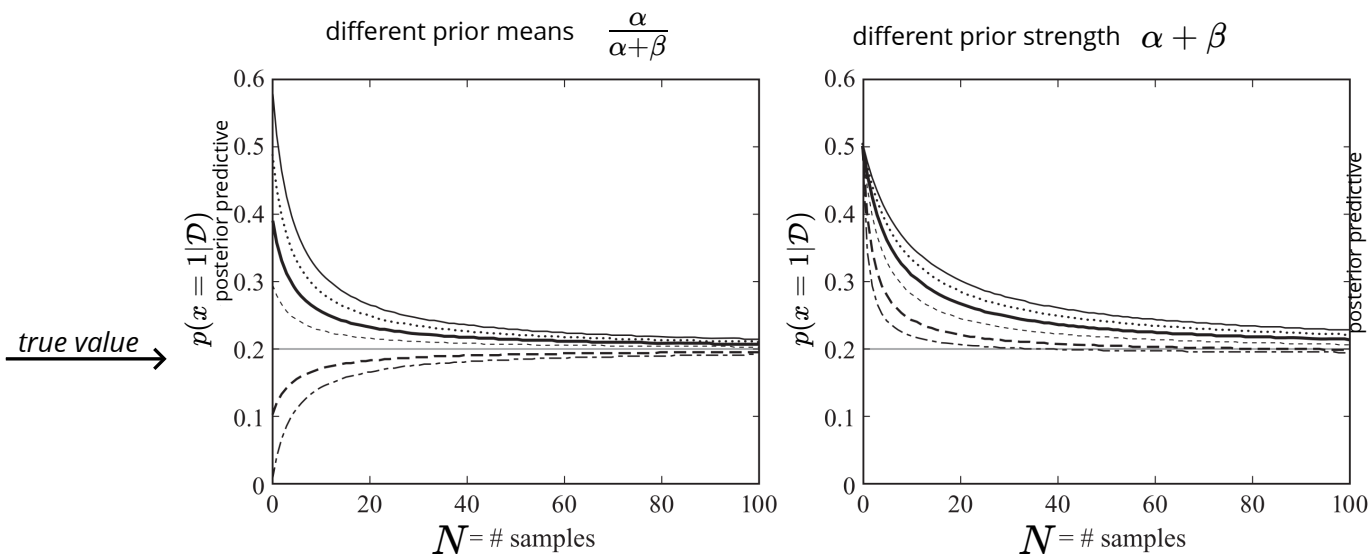
if we assume a uniform prior, the posterior predictive is $p(x = 1|\mathcal{D}) = \frac{N_h + 1}{N + 2}$ Laplace smoothing

Strength of the prior

with a **strong prior** we need many samples to really change the posterior

for Beta distribution $\alpha + \beta$ decides how strong the prior is

example as our dataset grows our estimate becomes more accurate



example: Koller & Friedman

Maximum a Posteriori (MAP)

sometimes it is difficult to work with the posterior dist. over parameters $p(\theta|\mathcal{D})$

alternative: use the parameter with the highest posterior probability

$$\theta^{MAP} = \arg\max_\theta p(\theta|\mathcal{D}) = \arg\max_\theta p(\theta)p(\mathcal{D}|\theta)$$ MAP estimate

compare with max-likelihood estimate *(the only difference is in the prior term)*

$$\theta^{MLE} = \arg\max_\theta p(\mathcal{D}|\theta)$$

example for the posterior $p(\theta|\mathcal{D}) = \text{Beta}(\theta|\alpha + N_h, \beta + N_t)$

MAP estimate is the **mode** of posterior $\theta^{MAP} = \frac{\alpha+N_h-1}{\alpha+\beta+N_h+N_t-2}$

compare with MLE $\theta^{MLE} = \frac{N_h}{N_h+N_t}$

they are equal for uniform prior $\alpha = \beta = 1$

Categorical distribution

what if we have more than two categories (e.g., loaded dice instead of coin)

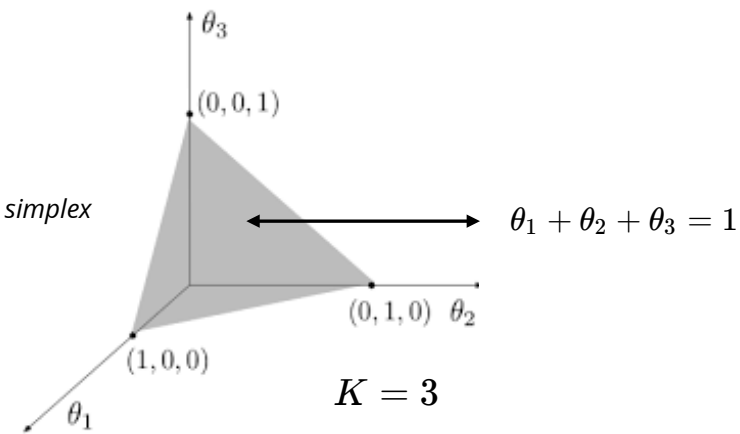
instead of Bernoulli we have multinoulli or **categorical** dist.

categories

$$\mathrm{Cat}(x|\theta) = \prod_{k=1}^{K} \theta_k^{\mathbb{I}(x=k)}$$

where $\sum_k \theta_k = 1$

θ *belongs to probability simplex*



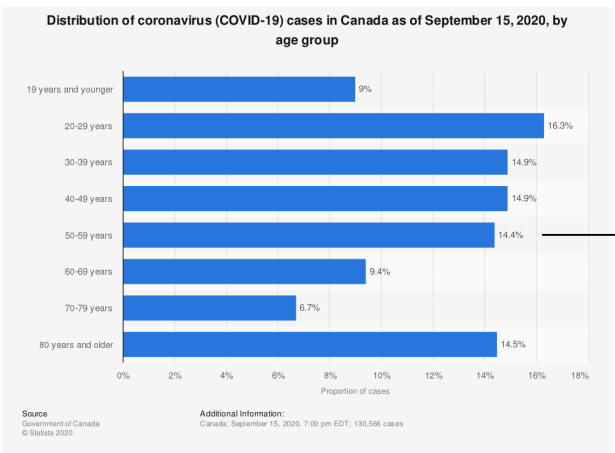
Maximum likelihood for categorical dist.

likelihood $p(\mathcal{D}|\theta) = \prod_{x \in \mathcal{D}} \mathrm{Cat}(x|\theta)$

log-likelihood $\ell(\theta, \mathcal{D}) = \sum_{x \in \mathcal{D}} \sum_k \mathbb{I}(x = k) \log(\theta_k)$

we need to solve $\frac{\partial}{\partial \theta_k} \ell(\theta, \mathcal{D}) = 0$ subject to $\sum_k \theta_k = 1$

similar to the binary case, max-likelihood estimate is given by data-frequencies $\theta_k{}^{MLE} = \frac{N_k}{N}$



example

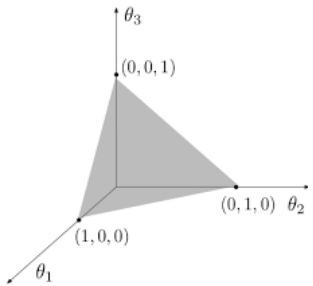
categorical distribution with K=8

frequencies are max-likelihood parameter estimates

$\theta_5^{MLE} = .149$

Dirichlet distribution

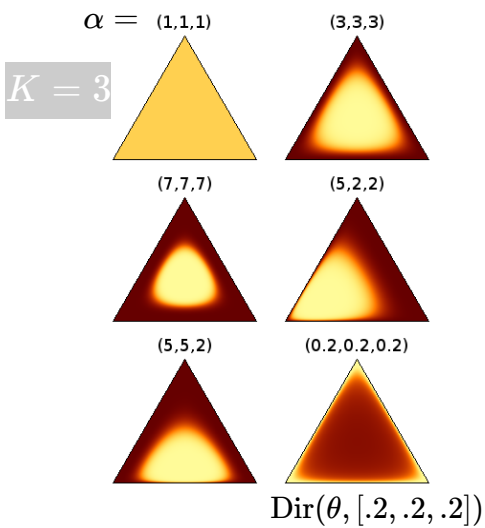
optional



is a distribution over the parameters θ of a Categorical dist.

is a generalization of Beta distribution to K categories

this should be a dist. over prob. simplex $\sum_k \theta_k = 1$



$$\mathrm{Dir}(\theta|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$$

normalization constant

vector of psedo-counts for K categories *(aka concentration parameters)*

$\alpha_k > 0 \; \forall k$

for $\alpha = [1, \ldots, 1]$, we get uniform distribution

for K=2, it reduces to Beta distribution

Dirichlet-Categorical conjugate pair

optional

Dirichlet dist. $\mathrm{Dir}(\theta|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$ is a conjugate prior for

Categorical dist. $\mathrm{Cat}(x|\theta) = \prod_k \theta_k^{\mathbb{I}(x=k)}$

prior $p(\theta) = \mathrm{Dir}(\theta|\alpha) \propto \prod_k \theta_k^{\alpha_k - 1}$

likelihood $p(\mathcal{D}|\theta) = \prod_k \theta_k^{N_k}$ we observe $\overset{\eta}{N_1, \ldots, N_K}$ values from each category

posterior $p(\theta|\mathcal{D}) = \mathrm{Dir}(\theta|\alpha + \eta) \propto \prod_k \theta_k^{N_k + \alpha_k - 1}$ again, we add the real counts to pseudo-counts

posterior predictive $p(x = k|\mathcal{D}) = \frac{\alpha_k + N_k}{\sum_{k'} \alpha_{k'} + N_{k'}}$

MAP $\theta_k^{MAP} = \frac{\alpha_k + N_k - 1}{(\sum_{k'} \alpha_{k'} + N_{k'}) - K}$

Summary

in ML we often build a probabilistic model of the data $p(x;\theta)$

learning a good model could mean **maximizing the likelihood** of the data

$$\max_\theta \log p(\mathcal{D}|\theta)$$ | sometimes closed form solution
for more complex p, we use numerical methods

an alternative is a **Bayesian approach**:

- maintain a **distribution** over model parameters
- can specify our **prior** knowledge $p(\theta)$
- we can use **Bayes rule** to update our belief after new oabservation $p(\theta|\mathcal{D})$
- we can make predictions using **posterior predictive** $p(x|\mathcal{D})$
- can be computationally **expensive** *(not in our examples so far)*

$$\max_\theta \log p(\mathcal{D}|\theta)p(\theta)$$

a middle path is **MAP estimate**:

- models our **prior** belief
- use a single point estimate and picks the model with highest posterior probability