10-301/601: Introduction to Machine Learning Lecture 9 – MLE & MAP

Henry Chai

6/3/24

Front Matter

- Announcements:
  - HW3 released 5/23, due 6/4 (tomorrow) at 11:59 PM
  - HW4 released 6/4 (tomorrow), due 6/11 at 11:59 PM
- Recommended Readings:
  - Mitchell, Estimating Probabilities

Probabilistic Learning

- Previously:
  - (Unknown) Target function, $c^*: \mathcal{X} \rightarrow \mathcal{Y}$
  - Classifier, $h : \mathcal{X} \rightarrow \mathcal{Y}$
  - Goal: find a classifier, h, that best approximates c^*
- Now:
  - (Unknown) Target *distribution*, $y \sim p^*(Y|\boldsymbol{x})$
  - Distribution, $p(Y|\boldsymbol{x})$
  - Goal: find a distribution, p, that best approximates p^*

Recall: $P(A \cap B) = P(A)P(B)$ if A & B are independent

Likelihood

- Given N independent, identically distribution (iid) samples $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$ of a random variable X
  - If X is discrete with probability mass function (pmf) $\underline{p}(X|\theta)$, then the *likelihood* of $\mathcal{D}$ is

$$\underline{L(\theta)} = \prod_{n=1}^{N} \underline{p\left(x^{(n)}|\theta\right)}$$

  - If X is continuous with probability density function (pdf) $f(X|\theta)$, then the *likelihood* of $\mathcal{D}$ is

$$L(\theta) = \prod_{n=1}^{N} f\left(x^{(n)}|\theta\right)$$

Log-Likelihood

- Given N independent, identically distribution (iid) samples $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$ of a random variable X
  - If X is discrete with probability mass function (pmf) $p(X|\theta)$, then the *log-likelihood* of $\mathcal{D}$ is

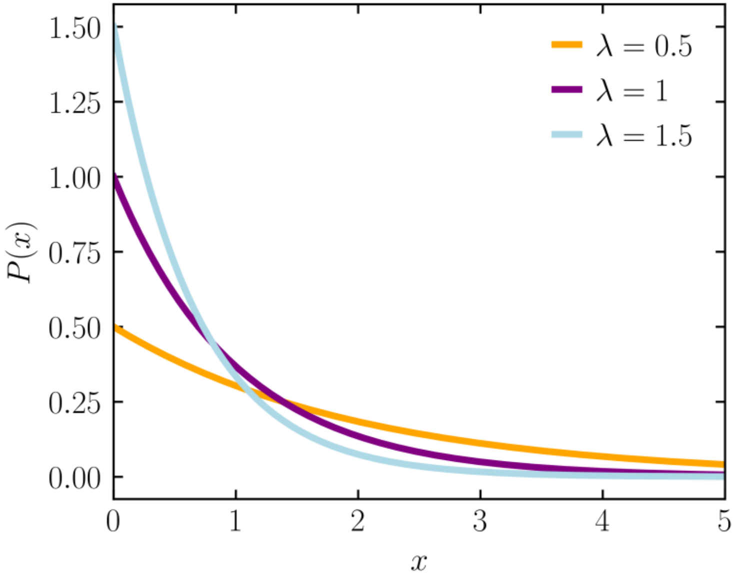
$$\underline{\ell}(\theta) = \log \prod_{n=1}^{N} p(x^{(n)}|\theta) = \sum_{n=1}^{N} \log p(x^{(n)}|\theta)$$

  - If X is continuous with probability density function (pdf) $f(X|\theta)$, then the *log-likelihood* of $\mathcal{D}$ is

$$\ell(\theta) = \log \prod_{n=1}^{N} f(x^{(n)}|\theta) = \sum_{n=1}^{N} \log f(x^{(n)}|\theta)$$

Maximum Likelihood Estimation (MLE)

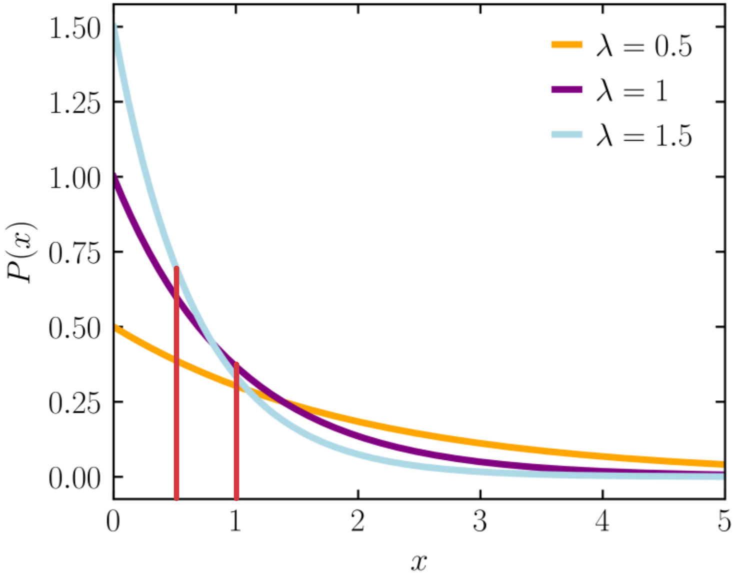
- Insight: every valid probability distribution has a finite amount of probability mass as it must sum/integrate to 1
- Idea: set the parameter(s) so that the likelihood of the samples is maximized
- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved data*
- Example: the exponential distribution



Source: https://en.wikipedia.org/wiki/Exponential_distribution#/media/File:Exponential_probability_density.svg

Maximum Likelihood Estimation (MLE)

- Insight: every valid probability distribution has a finite amount of probability mass as it must sum/integrate to 1
- Idea: set the parameter(s) so that the likelihood of the samples is maximized
- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved data*
- Example: the exponential distribution

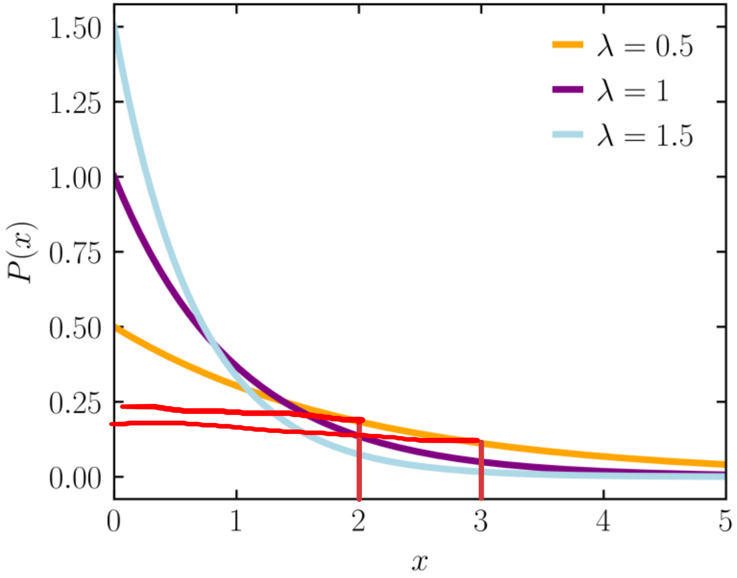


$\{x^{(1)} = 0.5,$
$x^{(2)} = 1\}$

Source: https://en.wikipedia.org/wiki/Exponential_distribution#/media/File:Exponential_probability_density.svg

Maximum Likelihood Estimation (MLE)

- Insight: every valid probability distribution has a finite amount of probability mass as it must sum/integrate to 1
- Idea: set the parameter(s) so that the likelihood of the samples is maximized
- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved data*
- Example: the exponential distribution



$\{x^{(1)} = 2,$
$x^{(2)} = 3\}$

Source: https://en.wikipedia.org/wiki/Exponential_distribution#/media/File:Exponential_probability_density.svg

General Recipe for Machine Learning

- Define a model and model parameters

- Write down an objective function

- Optimize the objective w.r.t. the model parameters

Recipe for MLE

- Define a model and model parameters

  - Specify the generative distribution along with the tunable parameters

- Write down an objective function

  - Maximize the log-likelihood of the data D

$$\ell_D(\theta) = \sum_{n=1}^{N} \log p(x^{(n)} | \theta)$$

- Optimize the objective w.r.t. the model parameters

  - Solve for θ in closed-form: take partial derivatives, set equal to 0 and solve.

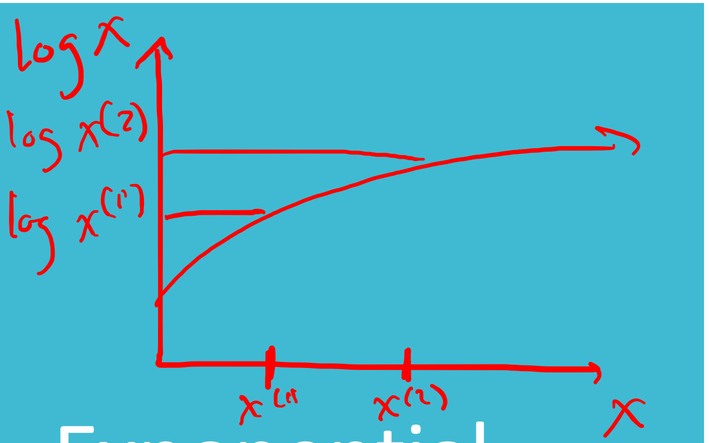
Exponential Distribution MLE

- The pdf of the exponential distribution is

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

- Given N iid samples $\{x^{(1)}, \dots, x^{(N)}\}$, the likelihood is

$$\mathcal{L}_{\mathcal{D}}(\lambda) = \prod_{n=1}^{N} f(x^{(n)}|\lambda) = \prod_{n=1}^{N} \lambda e^{-\lambda x^{(n)}}$$



Exponential Distribution MLE

- The pdf of the exponential distribution is

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

- Given N iid samples $\{x^{(1)}, \ldots, x^{(N)}\}$, the log-likelihood is

$$\ell_D(\lambda) = \sum_{n=1}^{N} \log \lambda e^{-\lambda x^{(n)}}$$

$$= \sum_{n=1}^{N} \left(\log \lambda + \log e^{-\lambda x^{(n)}}\right)$$

$$= N \log \lambda + \sum_{n=1}^{N} \left(-\lambda x^{(n)}\right)$$

$$\frac{\partial \ell_D}{\partial \lambda} = \frac{N}{\lambda} - \sum_{n=1}^{N} x^{(n)}$$

$$\Rightarrow \frac{N}{\hat{\lambda}} - \sum_{n=1}^{N} x^{(n)} = 0 \Rightarrow \frac{N}{\hat{\lambda}} = \sum_{n=1}^{N} x^{(n)} \Rightarrow \hat{\lambda} = \frac{N}{\sum_{n=1}^{N} x^{(n)}}$$

Bernoulli Distribution MLE

- A Bernoulli random variable takes value 1 with probability ϕ and value 0 with probability $1-\phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^{x}(1-\phi)^{1-x}$$

Coin Flipping MLE

$$\log a^b = b \log a$$

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1-\phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \underline{\phi}^{x}\underline{(1-\phi)}^{\boxed{1-x}}$$

Given $D = \{x^{(1)}, \ldots, x^{(N)}\}$ which we assume are i.i.d.

$$\ell_D(\phi) = \sum_{n=1}^{N} \log\left(\phi^{x^{(n)}}(1-\phi)^{1-x^{(n)}}\right)$$

$$= \sum_{n=1}^{N} \log \phi^{x^{(n)}} + \log(1-\phi)^{1-x^{(n)}}$$

$$= \sum_{n=1}^{N} x^{(n)} \log \phi + (1-x^{(n)}) \log(1-\phi)$$

$$= N_1 \log \phi + N_0 \log(1-\phi)$$

where N_i = the # of i's (0 or 1) in D

Coin Flipping MLE

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1 - \phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x (1 - \phi)^{1-x}$$

- The partial derivative of the log-likelihood is

if $\ell_D(\phi) = N_1 \log \phi + N_0 \log(1-\phi)$

then $\dfrac{\partial \ell_D}{\partial \phi} = \dfrac{N_1}{\phi} + \dfrac{N_0}{1-\phi}(-1) = \dfrac{N_1}{\phi} - \dfrac{N_0}{1-\phi}$

$\Rightarrow \dfrac{N_1}{\hat{\phi}} - \dfrac{N_0}{1-\hat{\phi}} = 0 \Rightarrow \dfrac{N_1}{\hat{\phi}} = \dfrac{N_0}{1-\hat{\phi}}$

$\Rightarrow N_1(1-\hat{\phi}) = N_0\hat{\phi} \Rightarrow N_1 = N_0\hat{\phi} + N_1\hat{\phi}$

$N_1 - N_1\hat{\phi} = N_0\hat{\phi} \Rightarrow$

$\Rightarrow \hat{\phi} = \dfrac{N_1}{N_0 + N_1} = \dfrac{N_1}{N}$

Given the result of your 5 coin flips, what is the MLE of ϕ for your coin?

0/5	0%
1/5	0%
2/5	0%
3/5	0%
4/5	0%
5/5	0%

Maximum a Posteriori (MAP) Estimation

- Insight: sometimes we have *prior* information we want to incorporate into parameter estimation
- Idea: use Bayes rule to reason about the *posterior* distribution over the parameters

– MLE finds $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(D \mid \theta)$

– MAP finds $\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\ p(\theta \mid D)$

$$= \underset{\theta}{\operatorname{argmax}}\ \frac{p(D \mid \theta)\, p(\theta)}{p(D)}$$

$$\boxed{p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta}$$

$$= \underset{\theta}{\operatorname{argmax}}\ \underbrace{p(D \mid \theta)}_{\text{likelihood}}\ \underbrace{p(\theta)}_{\text{prior}}$$

Recipe for MAP

- Define a model and model parameters
  - Specifying a generative distribution and a prior over each parameter
  - Assume i.i.d. samples D
- Write down an objective function
  - maximize th log-posterior of D

$$\ell_D^{MAP}(\Theta) = \log\left(P(D|\Theta)\,P(\Theta)\right) = \log p(\Theta) + \sum_{n=1}^{N} \log p(x^{(n)}|\Theta)$$

- Optimize the objective w.r.t. the model parameters
  - Solve in closed-form

Coin Flipping MAP

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1 - \phi$
- The pmf of the Bernoulli distribution is

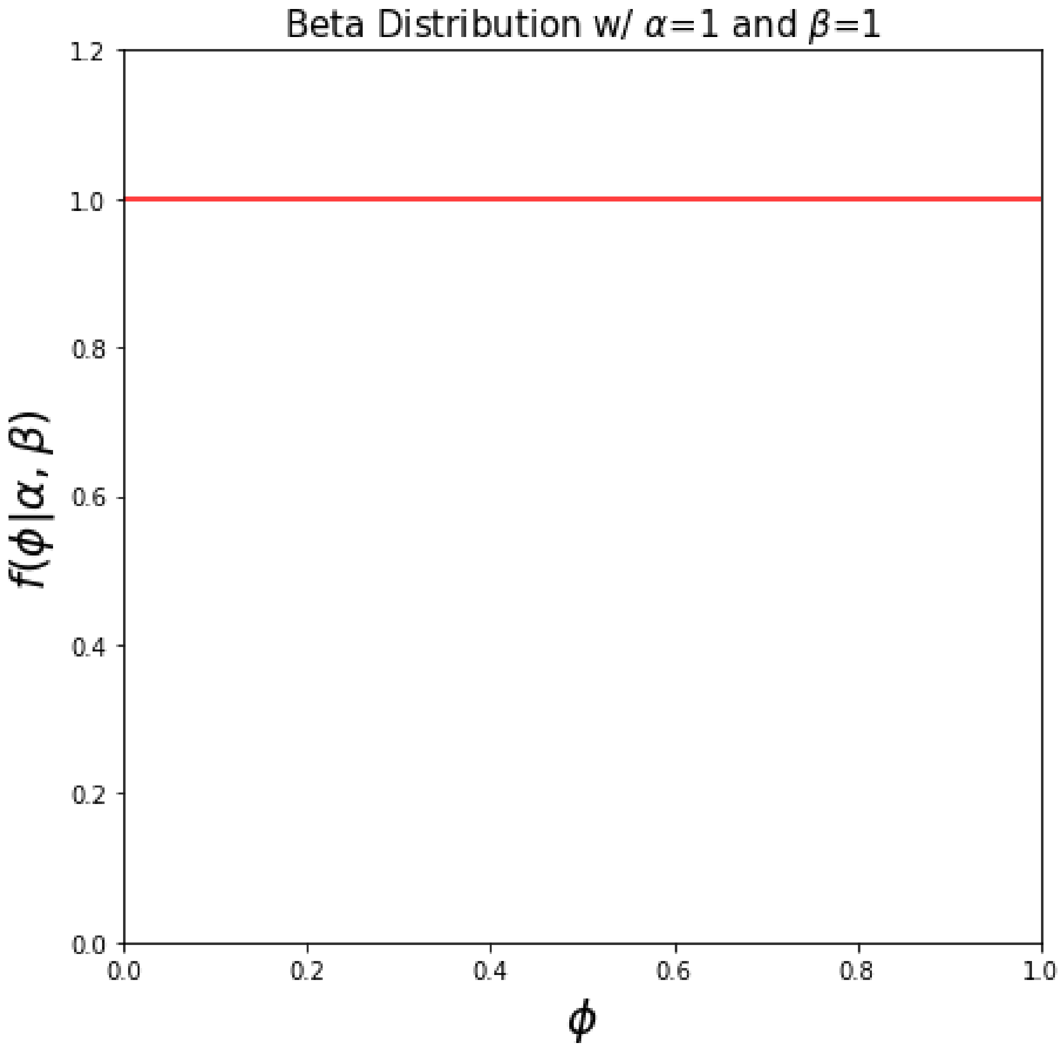
$$p(x|\phi) = \phi^{x}(1-\phi)^{1-x}$$

- Assume a Beta prior over the parameter ϕ, which has pdf

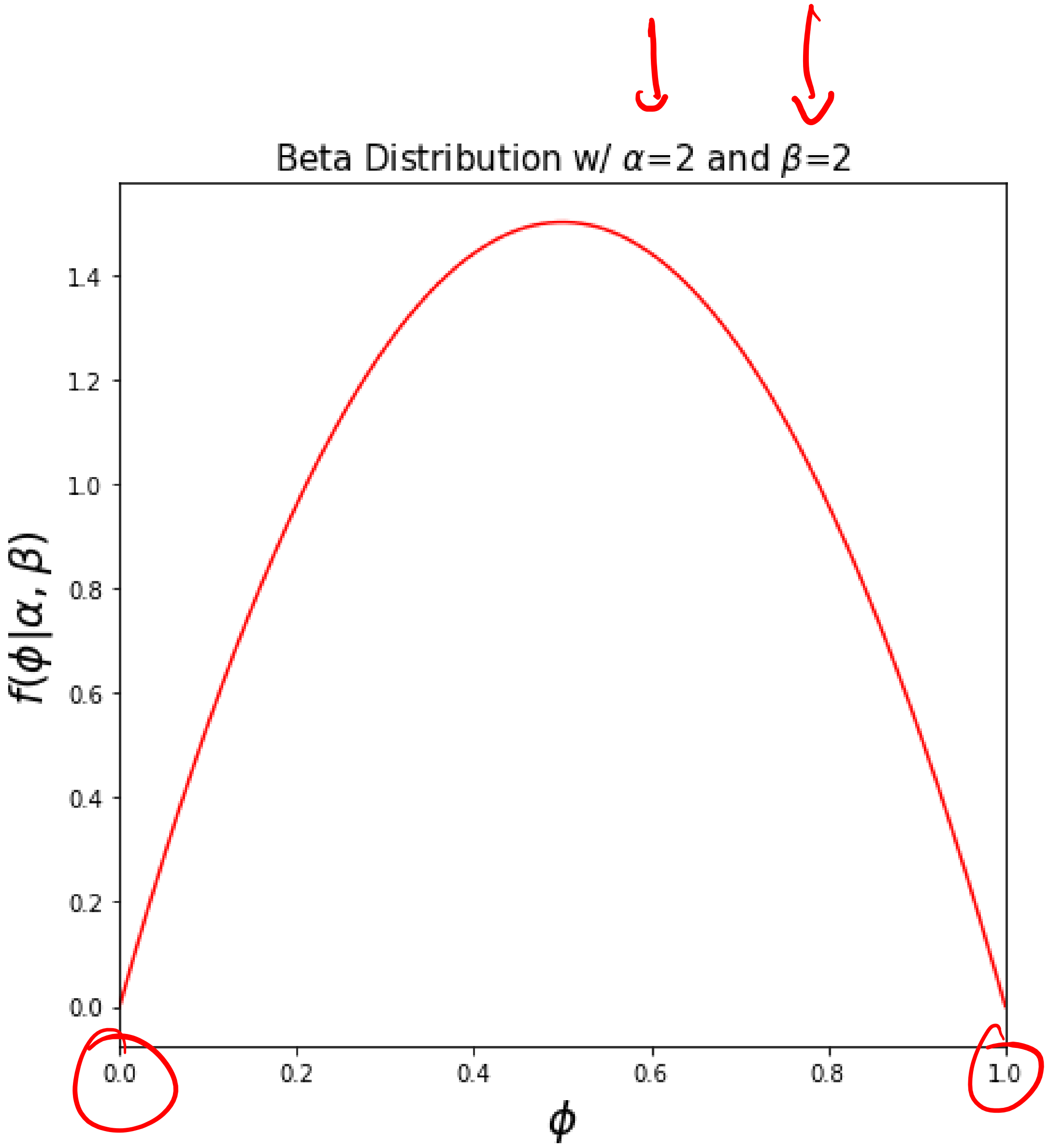
$$f(\phi|\alpha,\beta) = \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

where $\mathrm{B}(\alpha,\beta) = \int_0^1 \phi^{\alpha-1}(1-\phi)^{\beta-1}d\phi$ is a normalizing constant to ensure the distribution integrates to 1

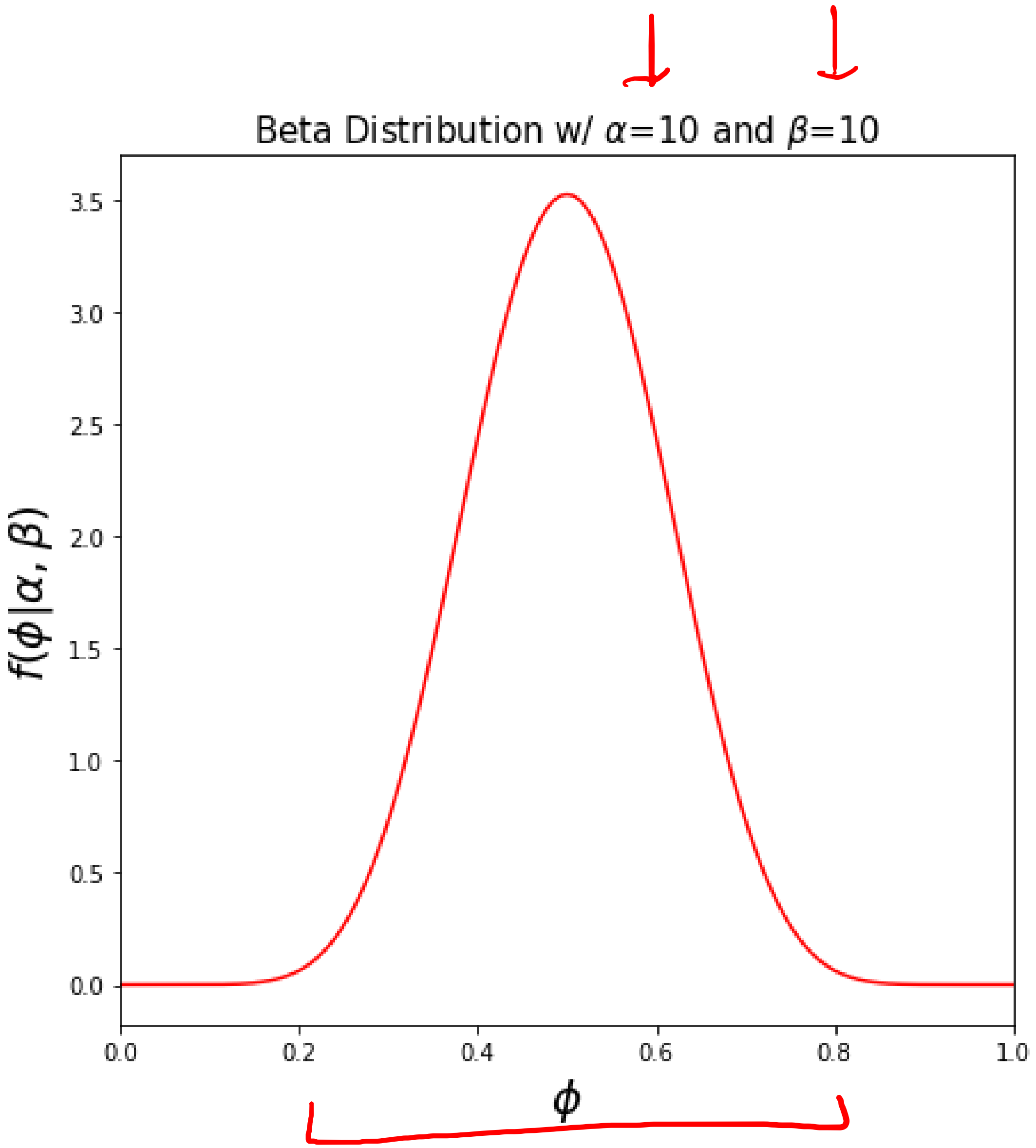
Beta Distribution



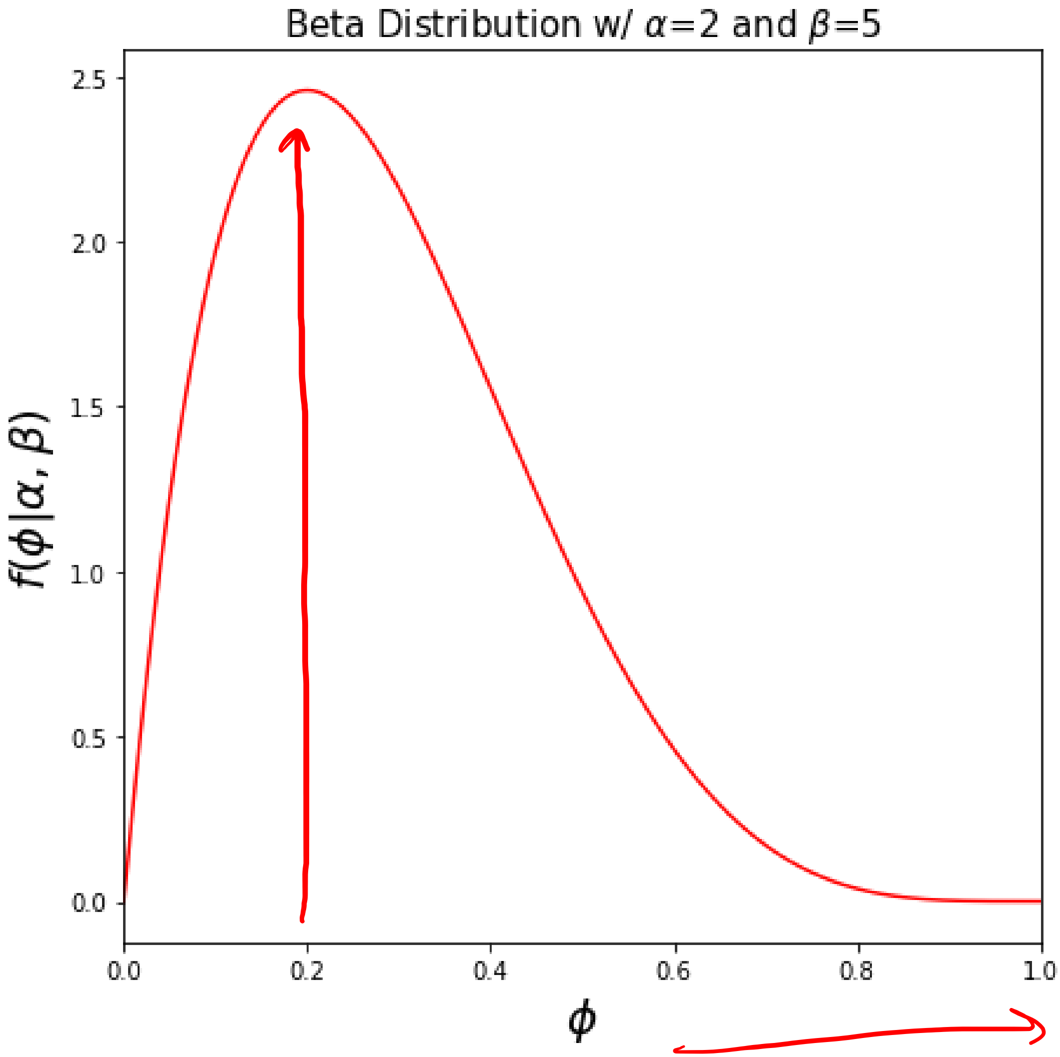
Beta Distribution



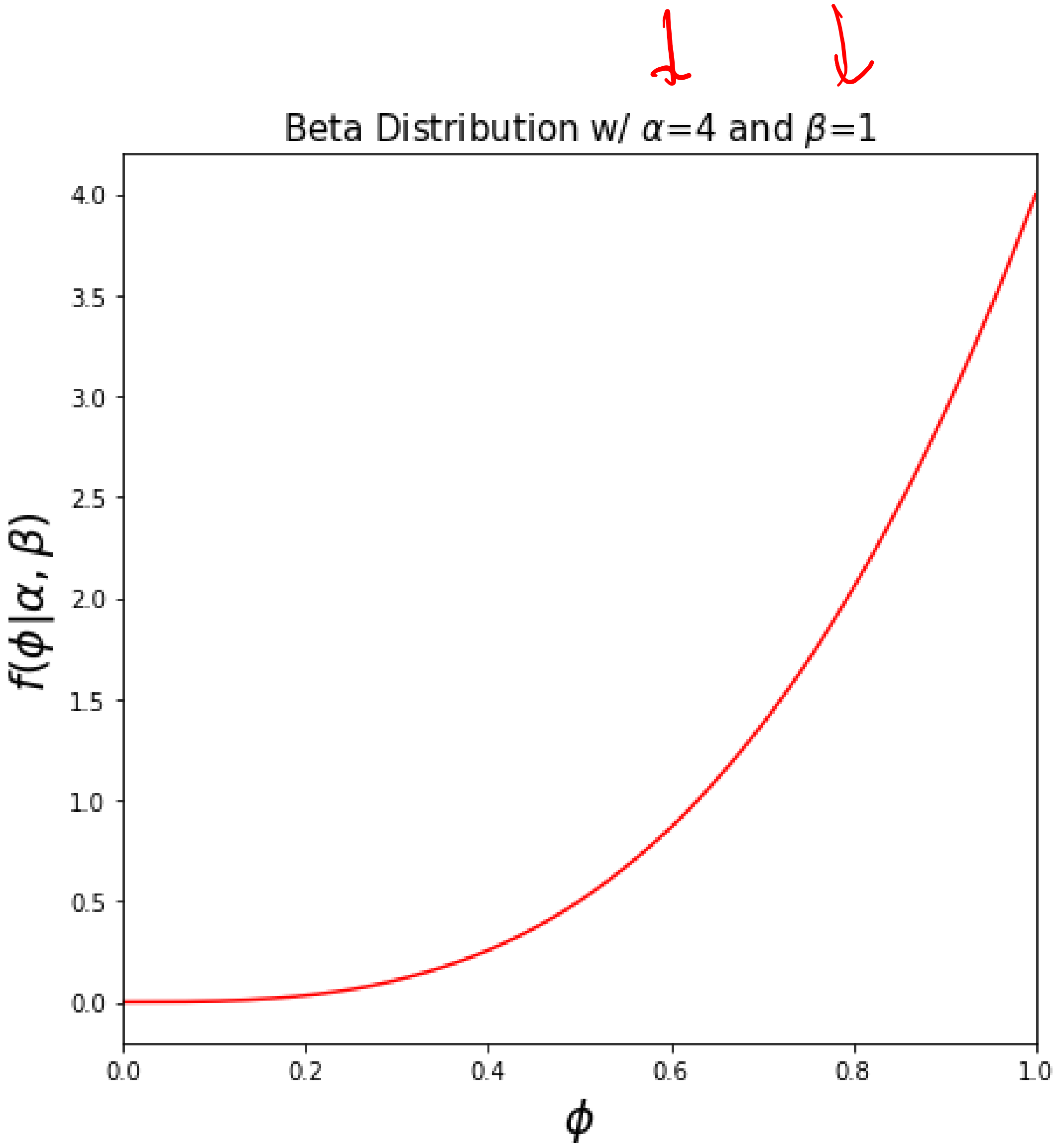
Beta Distribution



Beta Distribution



Beta Distribution



$$\log \frac{a}{b} = \log a - \log b$$

Coin Flipping MAP

- Given N iid samples $\{x^{(1)}, \dots, x^{(N)}\}$, the log-posterior is

$$\ell_D^{MAP}(\phi) = \log f(\phi \mid \alpha, \beta) + \sum_{n=1}^{N} \log p(x^{(n)} \mid \phi)$$

$$= \underbrace{\log \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{B(\alpha,\beta)}}_{} + \sum_{n=1}^{N} \log p(x^{(n)} \mid \phi)$$

$$= (\alpha-1)\log\phi + (\beta-1)\log(1-\phi) - \log B(\alpha,\beta)$$

$$+ N_1 \log\phi + N_0 \log(1-\phi)$$

$$= (N_1+\alpha-1)\log\phi + (N_0+\beta-1)\log(1-\phi)$$

$$- \log B(\alpha,\beta)$$

Coin Flipping MAP

- Given N iid samples $\{x^{(1)}, \dots, x^{(N)}\}$, the partial derivative of the log-posterior is

$$\frac{\partial \ell_D^{MAP}}{\partial \phi} = \frac{(N_1 + \alpha - 1)}{\phi} - \frac{(N_0 + \beta - 1)}{1 - \phi}$$

$$\vdots$$

$$\phi_{MAP} = \frac{N_1 + \alpha - 1}{(N_1 + \alpha - 1) + (N_0 + \beta - 1)}$$

$\alpha - 1$ and $\beta - 1$ are "pseudocounts" for the number of heads and tails that you've "previously observed"

Coin Flipping MAP: Example

- Suppose $\mathcal{D}$ consists of ten 1's or heads ($N_1 = 10$) and two 0's or tails ($N_0 = 2$):

$$\phi_{MLE} = \frac{10}{10+2} = \frac{10}{12}$$

- Using a Beta prior with $\alpha = 2$ and $\beta = 5$, then

Coin Flipping MAP: Example

- Suppose $\mathcal{D}$ consists of ten 1's or heads ($N_1 = 10$) and two 0's or tails ($N_0 = 2$):

$$\phi_{MLE} = \frac{10}{10+2} = \frac{10}{12}$$

- Using a Beta prior with $\alpha = 101$ and $\beta = 101$, then

Coin Flipping MAP: Example

- Suppose $\mathcal{D}$ consists of ten 1's or heads ($N_1 = 10$) and two 0's or tails ($N_0 = 2$):

$$\phi_{MLE} = \frac{10}{10+2} = \frac{10}{12}$$

- Using a Beta prior with $\alpha = 1$ and $\beta = 1$, then

Key Takeaways

- Probabilistic learning tries to learn a probability distribution as opposed to a classifier
- Two ways of estimating the parameters of a probability distribution given samples of a random variable:
  - Maximum likelihood estimation – maximize the (log-)likelihood of the observations
  - Maximum a posteriori estimation – maximize the (log-)posterior of the parameters conditioned on the observations
    - Requires a prior distribution, drawn from background knowledge or domain expertise