10-301/601: Introduction to Machine Learning Lecture 8 – MLE & MAP

Henry Chai

5/19/25

Front Matter

- Announcements:
  - HW2 released on 5/16, due 5/20 (tomorrow!) at 11:59 PM
  - HW3 to be released on 5/20 (tomorrow!), due 5/23 at 11:59 PM

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^*

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 *likelihood* of $\mathcal{D}$ is

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

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

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

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

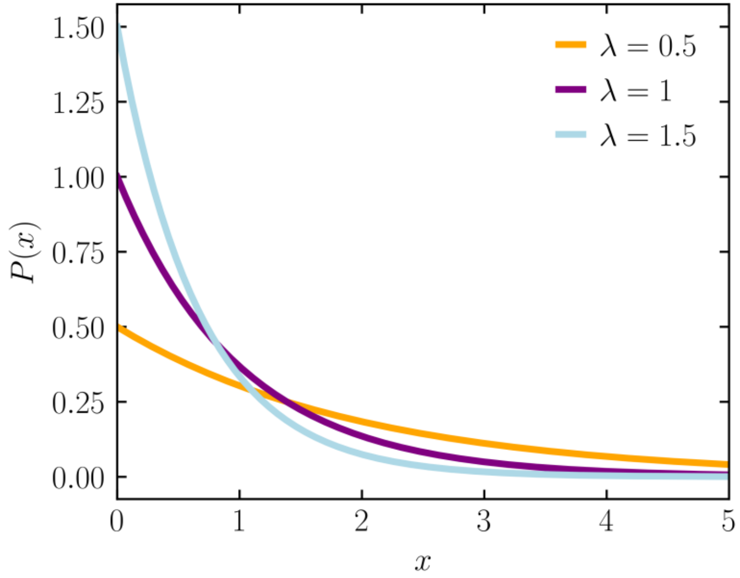
$$\ell(\theta) = \log \prod_{n=1}^{N} p\left(x^{(n)}|\theta\right) = \sum_{n=1}^{N} \log p\left(x^{(n)}|\theta\right)$$

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

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

Maximum Likelihood Estimation (MLE)

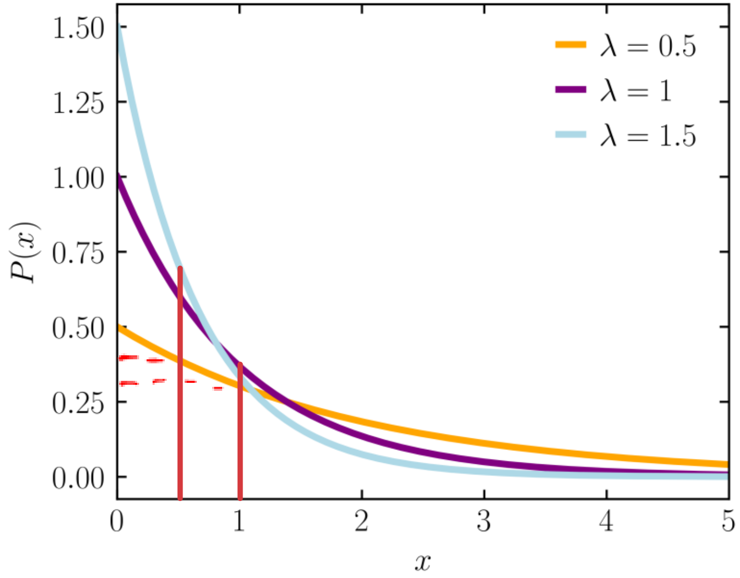
- Insight: every valid probability distribution has a finite amount of probability mass as it must sum/integrate to 1
- Idea: set the parameter(s) so that the likelihood of the samples is maximized
- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved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

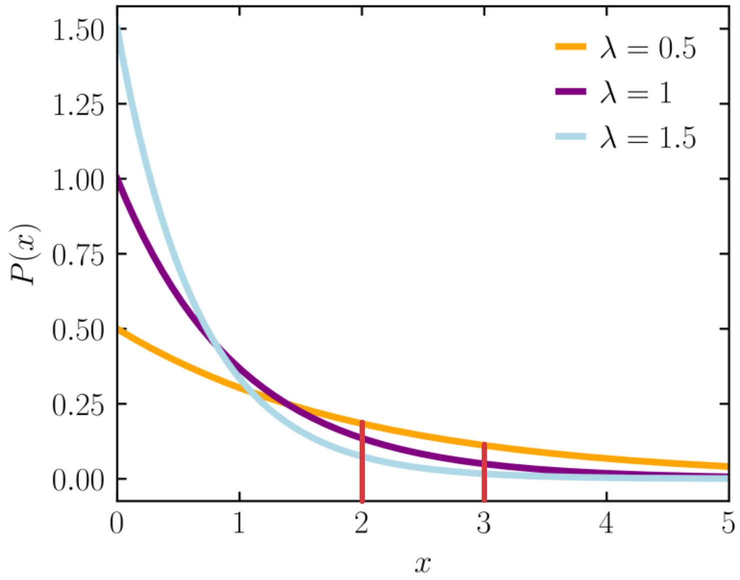


$\{x^{(1)} = 0.5, x^{(2)} = 1\}$

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 story" of D
  i.e. pick the distribution we're going to fit

- Write down an objective function

  – maximize the log-likelihood of D as a function of Θ

  $$\ell(\Theta) = \sum_{n=1}^{N} \log p(\vec{x}^{(n)} \mid \Theta)$$

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

  Solve "in closed form" using the critical point method

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

$$L(\lambda) = \prod_{n=1}^{N} \lambda e^{-\lambda x^{(n)}}$$

$$\log(a \cdot bc) = \log a + \log b + \log c$$

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 log-likelihood is

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

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

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

$$\Rightarrow \frac{\partial \ell}{\partial \lambda} = N\left(\frac{1}{\lambda}\right) - \sum_{n=1}^{N} x^{(n)}$$

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

$$\Rightarrow \hat{\lambda} = N \Big/ \sum_{n=1}^{N} x^{(n)}$$

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

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

$\log(a^b c^d)$
$= \log a^b + \log c^d$
$= b \log a + d \log c$

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

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

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

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

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

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

where N_j = # of j's in my dataset $\{x^{(1)}, \ldots, x^{(N)}\}$

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

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

$$\Rightarrow \frac{\partial \ell}{\partial \phi} = \frac{N_1}{\phi} + \frac{N_0}{1-\phi}(-1)$$

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

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

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

Lecture 8 + 9 Polls

0 surveys completed

0 surveys underway

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

- 0/5
- 1/5
- 2/5
- 3/5
- 4/5
- 5/5

$p(\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \mid \theta)$

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 $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(\theta \mid D)$ (posterior)

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

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

likelihood: $p(D \mid \theta)$; prior: $p(\theta)$

$$= \underset{\theta}{\operatorname{argmax}}\ \log p(D \mid \theta) + \log p(\theta)$$

(log-posterior)

Recipe for MAP

- Define a model and model parameters

  – Specify a generative story w/ a prior over the parameters

- Write down an objective function

$$\ell_{MAP}(\Theta) = \sum_{n=1}^{N} \log p(x^{(n)} \mid \Theta) + \log p(\Theta)$$

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

  Solve in closed form using the critical point method

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

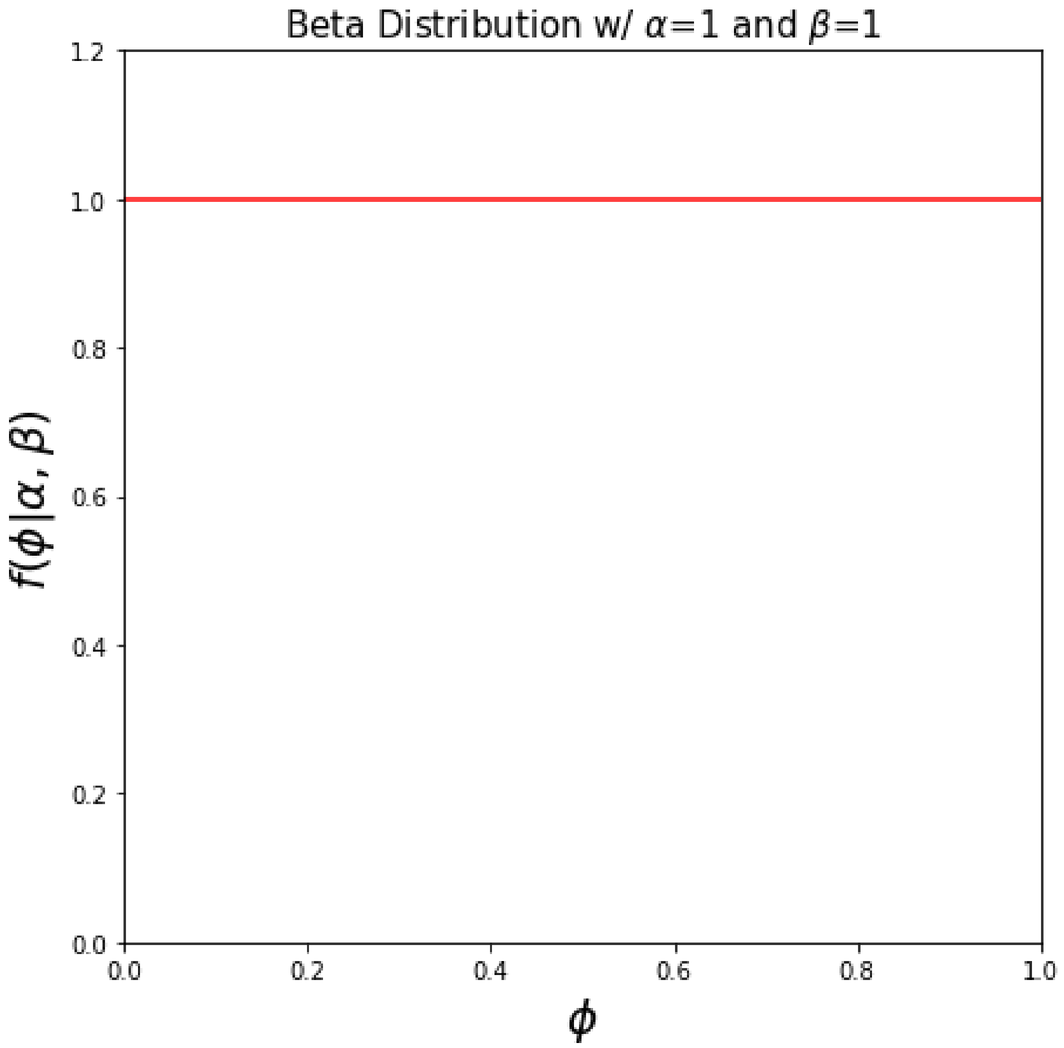
$$p(x|\phi) = \phi^x (1-\phi)^{1-x}$$

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

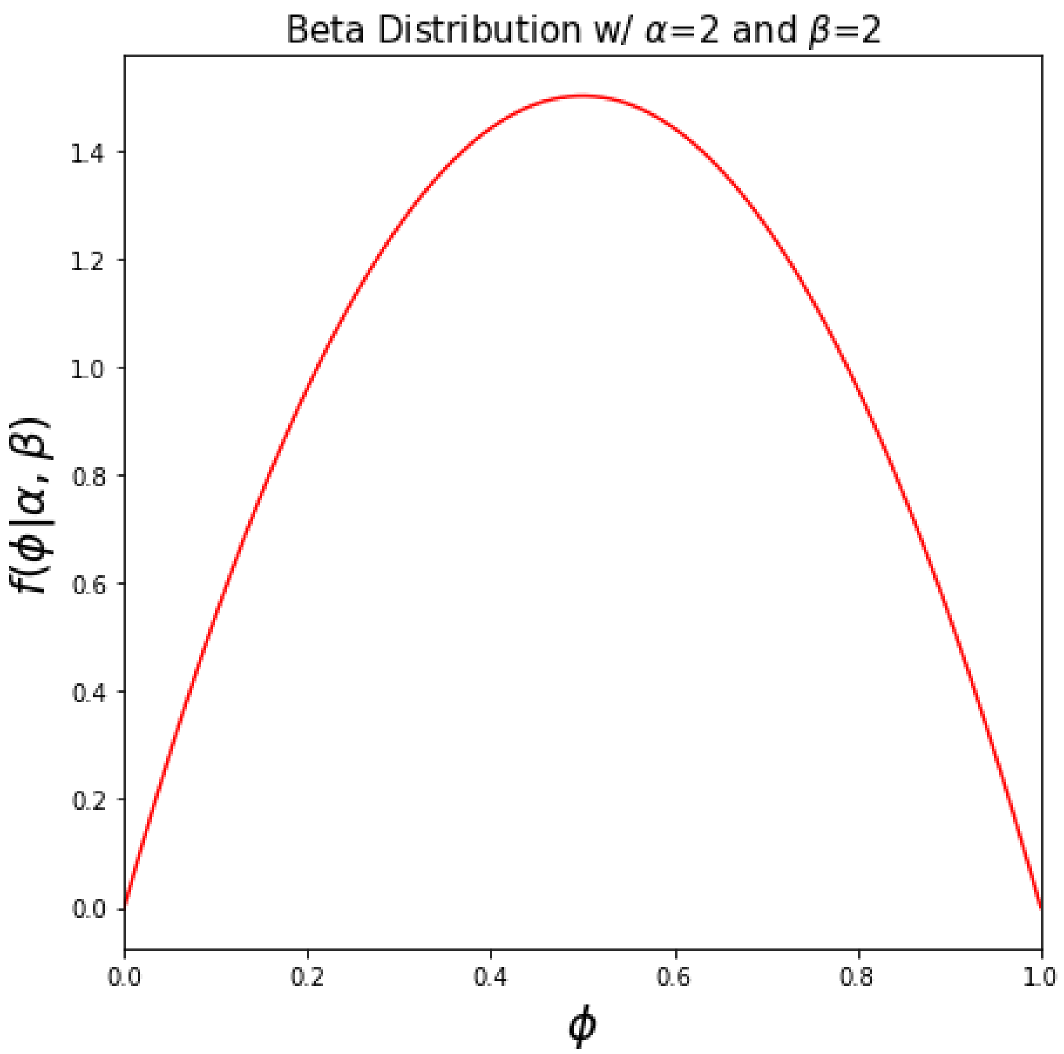
$$f(\phi|\alpha,\beta) = \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

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

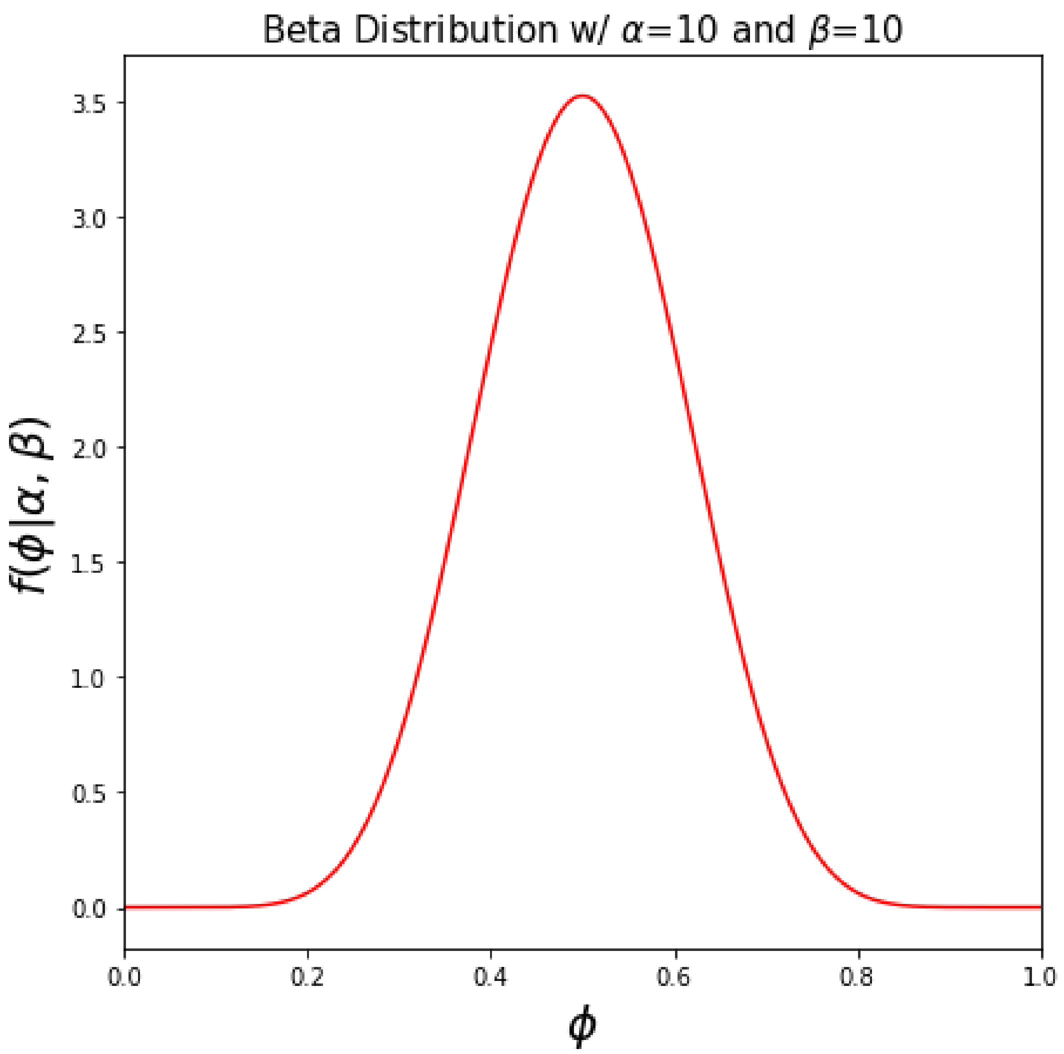
Beta Distribution



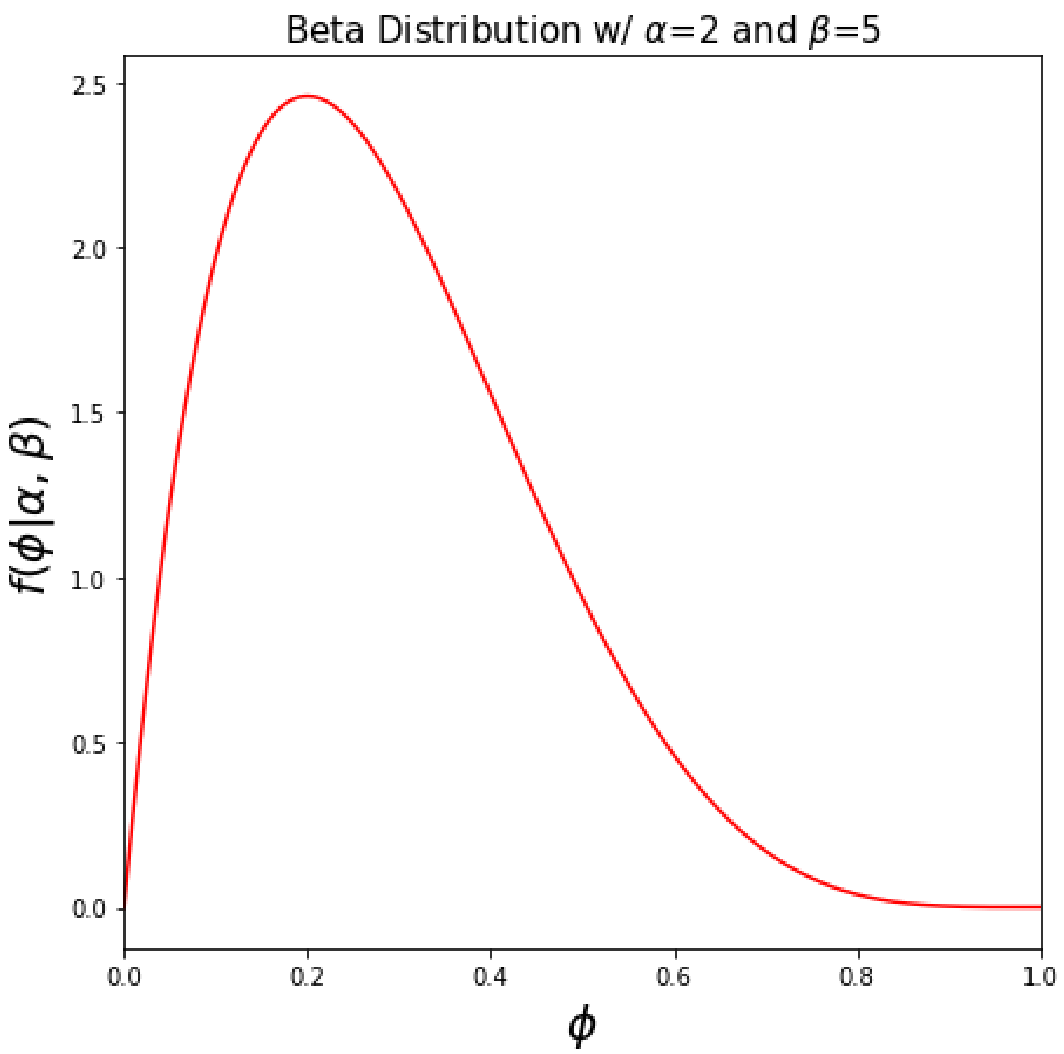
Beta Distribution



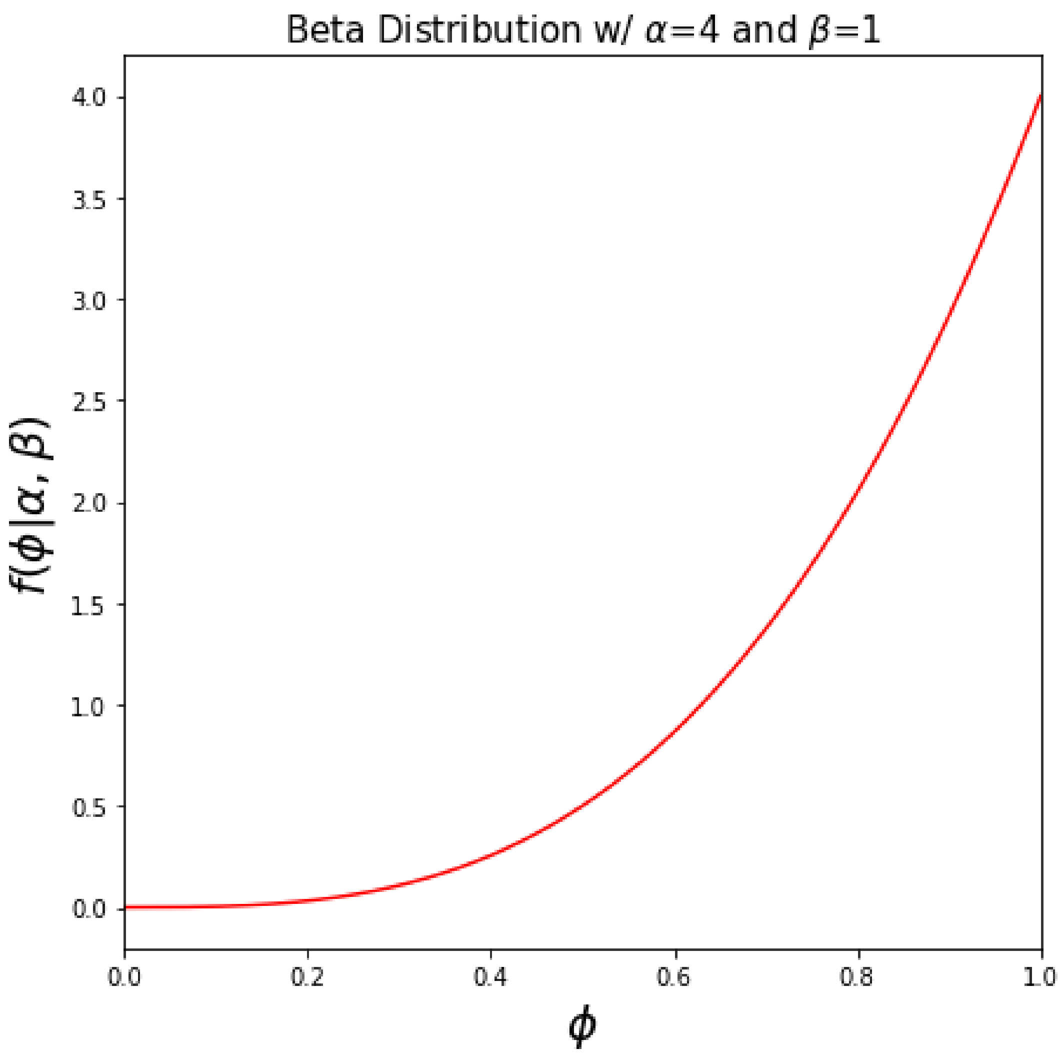
Beta Distribution



Beta Distribution



Beta Distribution



Coin Flipping MAP

$$\log\left(\frac{a}{b}\right) = \log a - \log b$$

$$\log(p(D|\theta))$$

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

$$\ell_{MAP}(\phi) = \sum_{n=1}^{N} \log p(x^{(n)} \mid \phi) + \log p(\phi)$$

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

$$\log\left(\frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{B(\alpha,\beta)}\right)$$

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

$$\ell_{MAP}(\phi) = (N_1+\alpha-1)\log\phi + (N_0+\beta-1)\log(1-\phi) - \log B(\alpha,\beta)$$

Coin Flipping MAP

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

$$\ell_{MAP}(\phi) = (N_1 + \alpha - 1)\log\phi + (N_0 + \beta - 1)\log(1-\phi) - \log B(\alpha, \beta)$$

$$\frac{\partial \ell_{MAP}}{\partial \phi} = \frac{N_1 + \alpha - 1}{\phi} - \frac{N_0 + \beta - 1}{1 - \phi}$$

$$\vdots$$

$$\hat{\phi} = \frac{N_1 + \alpha - 1}{(N_1 + \alpha - 1) + (N_0 + \beta - 1)}$$

$(\alpha - 1)$ & $(\beta - 1)$ are "pseudocounts" of heads & tails

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

$$\phi_{MAP} = \frac{110}{110+102} = \frac{110}{212} \approx 0.5$$

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