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

Henry Chai

6/5/23

Front Matter

- Announcements:
  - Quiz 3: Linear Regression & Optimization on 6/6 (tomorrow!)
- 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^*

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

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

P(A∩B) = P(A)P(B) if A and B are independent

  - 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 $\mathcal{D} = \left\{x^{(1)}, \ldots, x^{(N)}\right\}$ 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

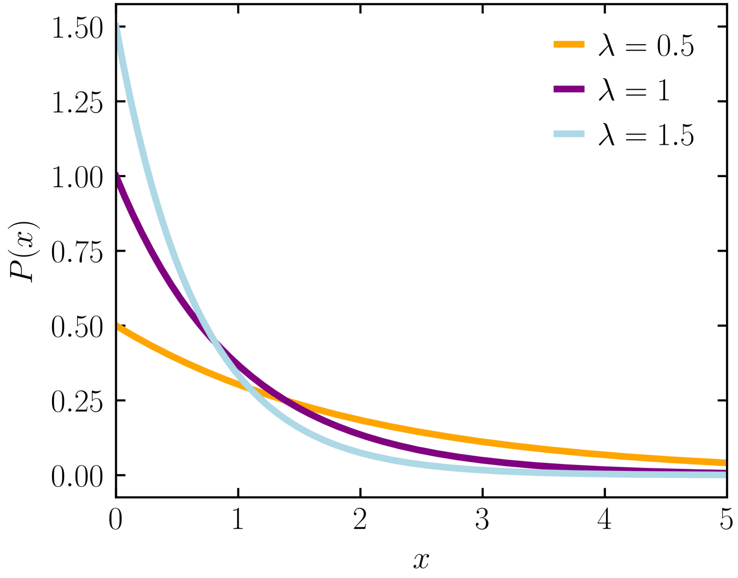
$$\ell(\theta) = \log \prod_{n=1}^{N} p\left(x^{(n)}|\theta\right) = \sum_{n=1}^{N} \log p\left(x^{(n)}|\theta\right)$$

  - 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)

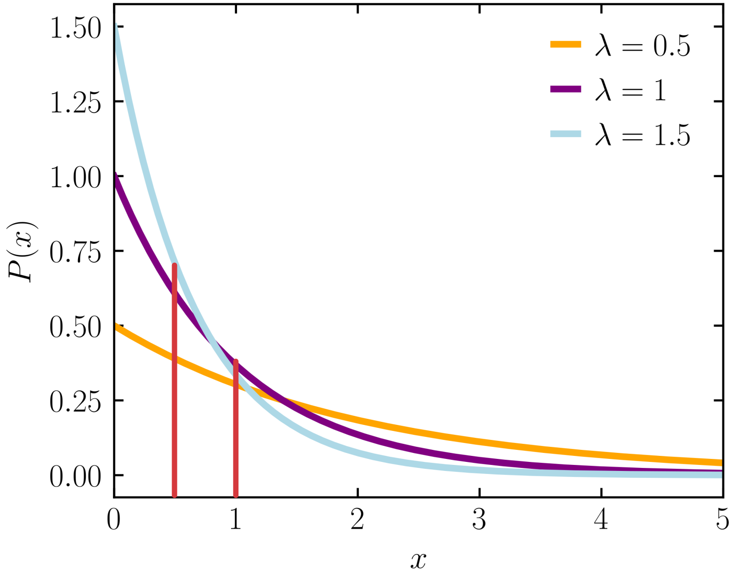
- Insight: every valid probability distribution has a finite amount of probability mass as it must sum/integrate to 1
- Idea: set the parameter(s) so that the likelihood of the samples is maximized
- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved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

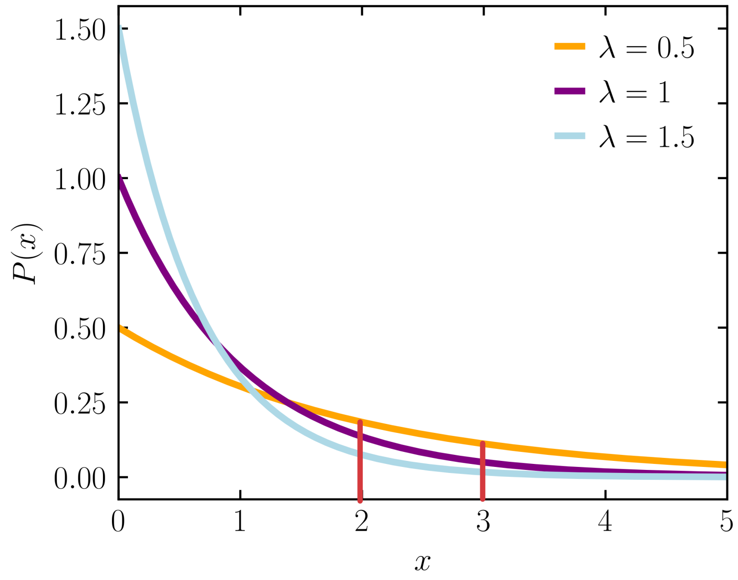


$$\{x^{(1)} = 0.5, x^{(2)} = 1\}$$

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 a "generative story" = pick a data-generating distribution
- Write down an objective function
  - Maximize the log-likelihood of $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$

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

- Optimize the objective w.r.t. the model parameters
  - Solve in closed-form by taking partial derivatives and setting them equal to 0 and solving

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 likelihood is

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

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

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

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

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

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

Exponential Distribution MLE

- The pdf of the exponential distribution is
$$f(x|\lambda) = \lambda e^{-\lambda x}$$

- Given N iid samples $\{x^{(1)}, \dots, x^{(N)}\}$, the log-likelihood is
$$\ell(\lambda) = \sum_{n=1}^{N} \log f\left(x^{(n)}|\lambda\right) = \sum_{n=1}^{N} \log \lambda e^{-\lambda x^{(n)}}$$
$$= \sum_{n=1}^{N} \log \lambda + \log e^{-\lambda x^{(n)}} = \underbrace{N \log \lambda - \lambda \sum_{n=1}^{N} x^{(n)}}$$

- Taking the partial derivative and setting it equal to 0 gives

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

$$\frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{N}{\lambda^2} \leq 0 \qquad \rightarrow \frac{N}{\hat{\lambda}} = \sum_{n=1}^{N} x^{(n)} \rightarrow \hat{\lambda} = \frac{N}{\sum_{n=1}^{N} x^{(n)}}$$

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

- 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 some observations $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$

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

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

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

where N_i is the # of i's in $\mathcal{D}$

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

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

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

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

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

$$\rightarrow N_1 = (N_1 + N_0)\hat{\phi} \rightarrow \hat{\phi} = \frac{N_1}{N_1 + N_0}$$

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

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

$$\rightarrow N_1\left(1-\hat{\phi}\right) = N_0\hat{\phi} \rightarrow N_1 = \hat{\phi}(N_0 + N_1)$$

$$\rightarrow \hat{\phi} = \frac{N_1}{N_0 + N_1}$$

- where N_1 is the number of 1's in $\left\{x^{(1)}, \ldots, x^{(N)}\right\}$ and N_0 is the number of 0's

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

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 $\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta) = \underset{\theta}{\operatorname{argmax}}\, p(D|\theta)$

likelihood

MAP: finds $\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|D)$

prior

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

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

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

Recipe for MAP

- Define a model and model parameters
  - Specify a generative story including the prior distribution (???)
- Write down an objective function
  - Maximize the log-posterior of $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$

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

- 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

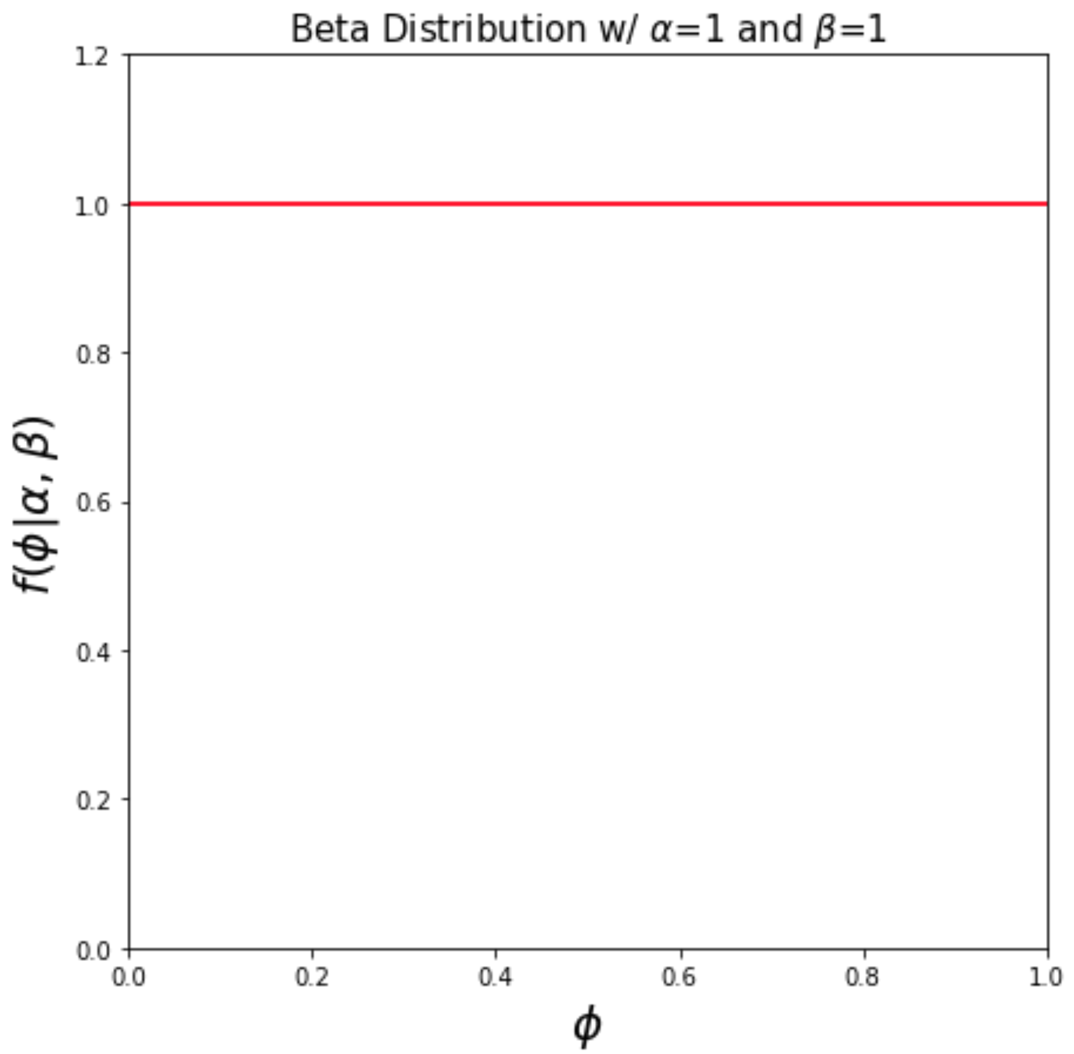
$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$

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

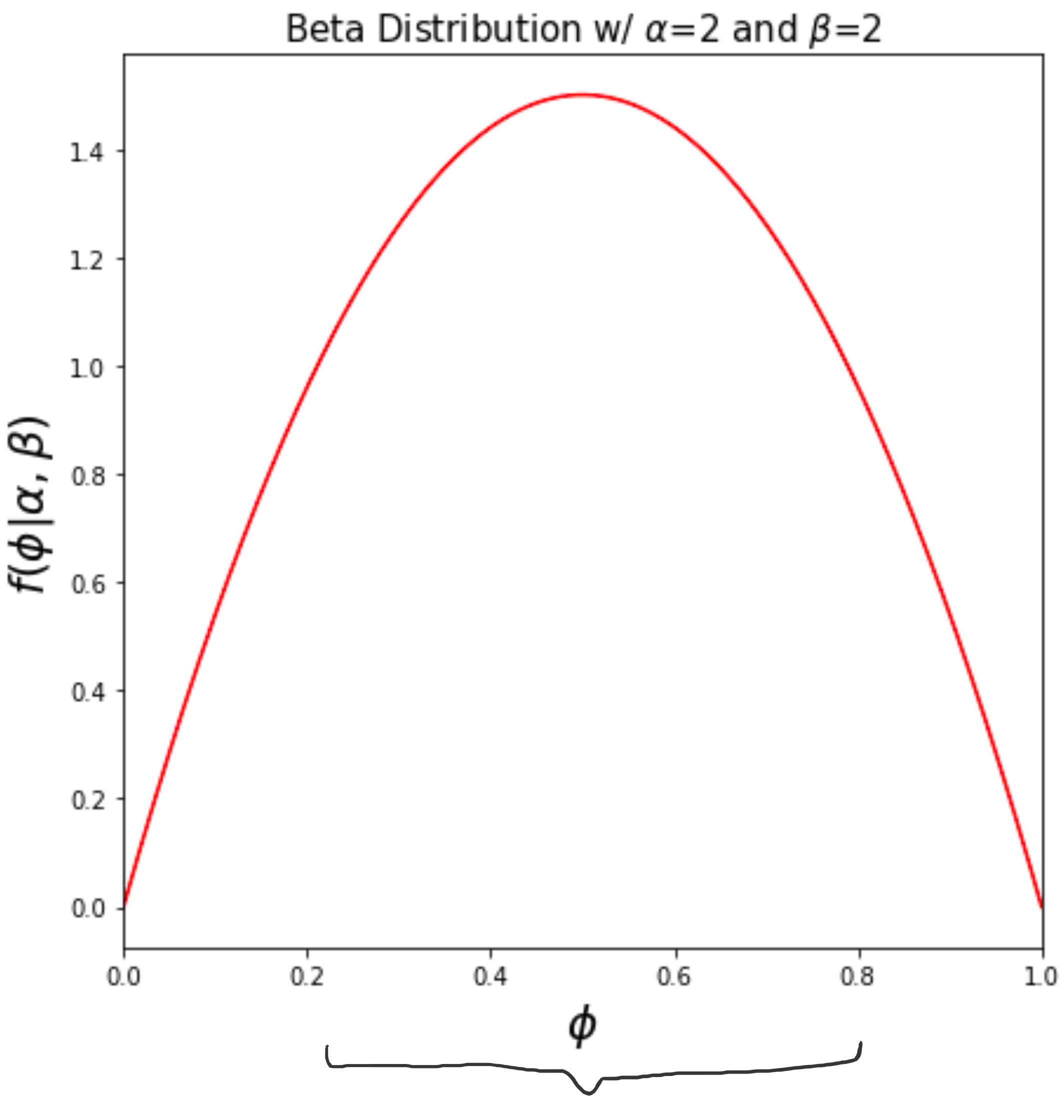
$$f(\phi|\alpha,\beta) = \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

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

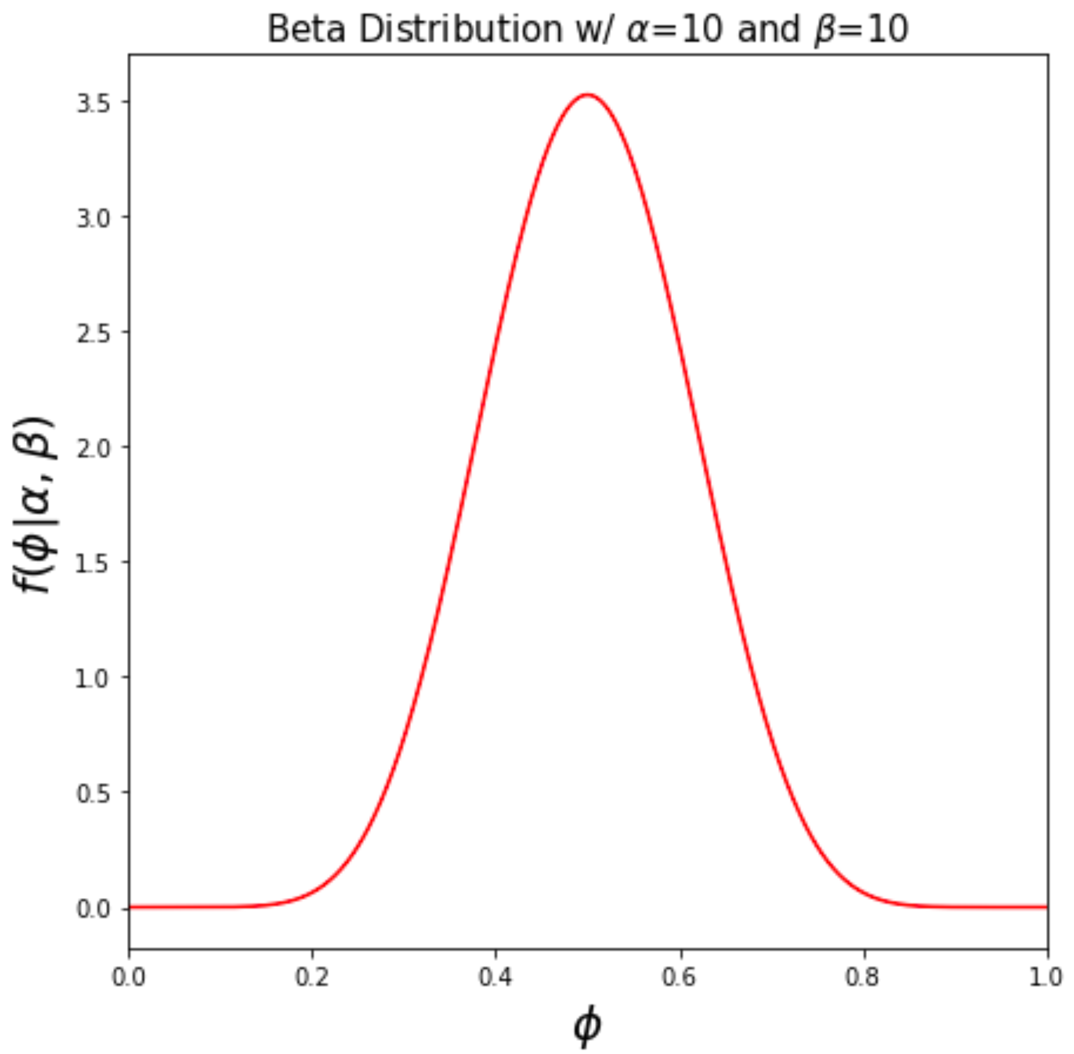
Beta Distribution



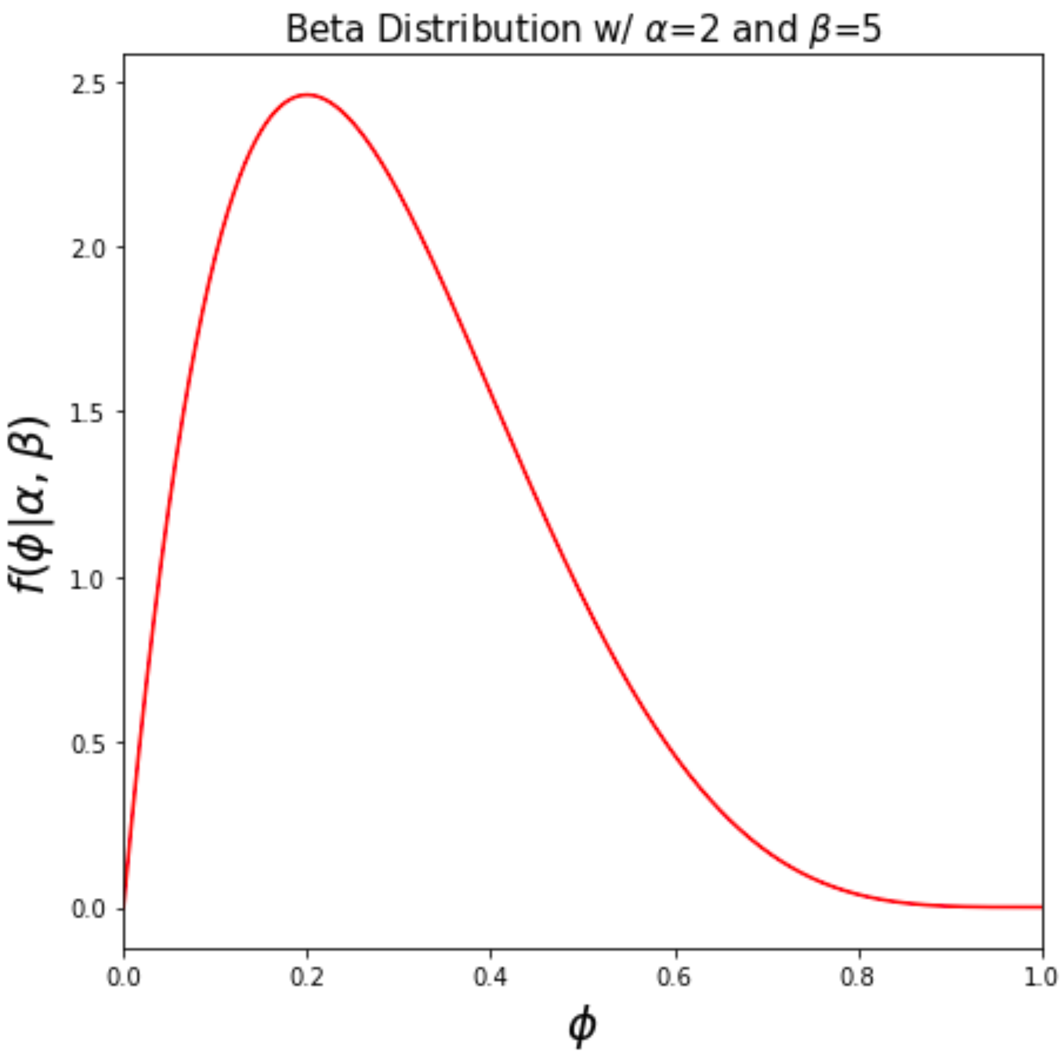
Beta Distribution



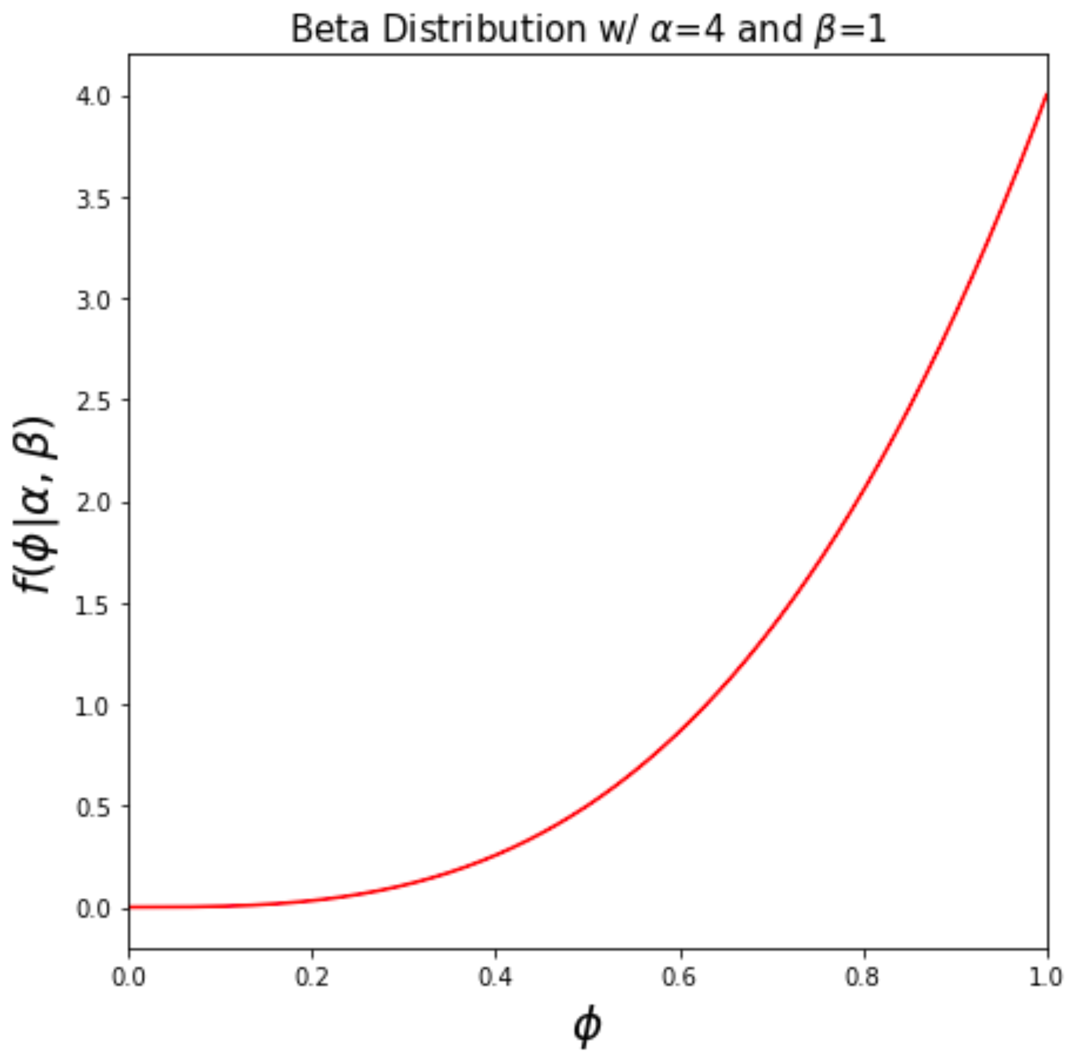
Beta Distribution



Beta Distribution



Beta Distribution



Coin Flipping MAP

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

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

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

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

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

Coin Flipping MAP

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

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

$$\vdots$$

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

$\alpha - 1$ is a "pseudocount" of the # of heads you've "previously observed"

$\beta - 1$ " " # of 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

$$\phi_{MAP} = \frac{(2-1+10)}{(2-1+10)+(5-1+2)} = \frac{11}{17} < \frac{10}{12}$$

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

$$\phi_{MAP} = \frac{(101-1+10)}{(101-1+10)+(101-1+2)} = \frac{110}{212} \approx \frac{1}{2}$$

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