CS 2750 Machine Learning
Lecture 5

Density estimation

Milos Hauskrecht
milos@pitt.edu
5329 Sennott Square

Density estimation

Density estimation: is an unsupervised learning problem

- **Goal:** Learn a model that represent the relations among attributes in the data

$$D = \{D_1, D_2, .., D_n\}$$

Data: $D_i = \mathbf{x}_i$ a vector of attribute values

Attributes:

- modeled by random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ with
  - **Continuous or discrete valued variables**

Density estimation: learn an underlying probability distribution model : $p(\mathbf{X}) = p(X_1, X_2, \ldots, X_d)$ **from D**

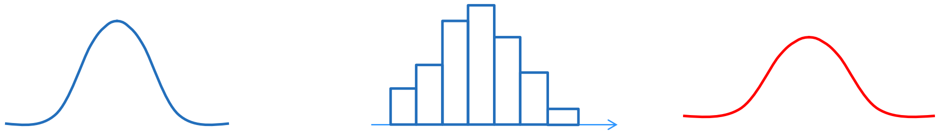
Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$

$D_i = \mathbf{x}_i$ a vector of attribute values

Objective: estimate the model of the underlying probability distribution over variables $\mathbf{X}$, $p(\mathbf{X})$, using examples in D

true distribution $p(\mathbf{X})$ → n samples $D = \{D_1, D_2, .., D_n\}$ → estimate $\hat{p}(\mathbf{X})$

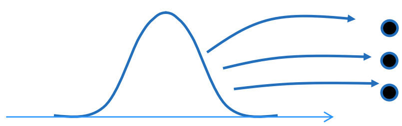


Density estimation

true distribution $p(\mathbf{X})$ → n samples $D = \{D_1, D_2, .., D_n\}$ → estimate $\hat{p}(\mathbf{X})$

Standard (iid) assumptions: Samples

- **are independent of each other**
- **come from the same (identical) distribution (fixed $p(\mathbf{X})$)**



Independently drawn instances from the same fixed distribution

Density estimation

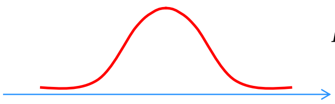
Types of density estimation:

(1) Parametric

- the distribution is modeled using a set of parameters Θ

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \Theta)$$

- **Estimation:** find parameters Θ fitting the data D
- **Example:** estimate the mean and covariance of a normal distribution



$$\hat{p}(x) = N(x \mid \mu, \sigma)$$

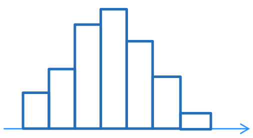
Density estimation

Types of density estimation:

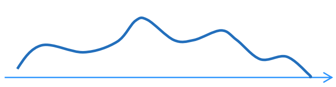
(2) Non-parametric

- The model of the distribution utilizes all examples in D
- As if all examples were parameters of the distribution
- $$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid D)$$
- **Examples:**

histogram



Kernel density estimation



Learning via parameter estimation

In this lecture we consider **parametric density estimation**

Basic settings:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$
- **A model of the distribution** over variables in ***X*** with parameters Θ : $\hat{p}(\mathbf{X} \mid \Theta)$
- **Data** $D = \{D_1, D_2, .., D_n\}$

- **Objective:** find parameters Θ such that $p(\mathbf{X} \mid \Theta)$ fits data D the best

- How to measure the goodness of fit or alternative the error?

ML Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\Theta})$ **Data** $D = \{D_1, D_2, .., D_n\}$

$\boxed{\max_{\Theta} p(D \mid \Theta, \xi)}$

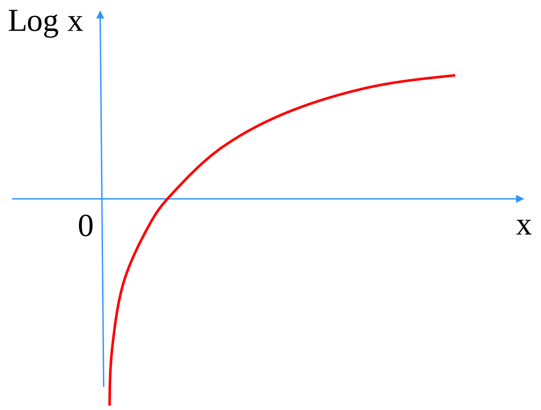
- **Maximum likelihood (ML)**
  - Find Θ that maximizes likelihood $p(D \mid \Theta, \xi)$

$$
\begin{aligned}
&P(D \mid \Theta, \xi) = P(D_1, D_2, .., D_n \mid \Theta, \xi) \\
&= P(D_1 \mid \Theta, \xi) P(D_2 \mid \Theta, \xi) \ldots P(D_n \mid \Theta, \xi) \\
&= \prod_{i=1}^{n} P(D_i \mid \Theta, \xi)
\end{aligned}
$$

Independent examples

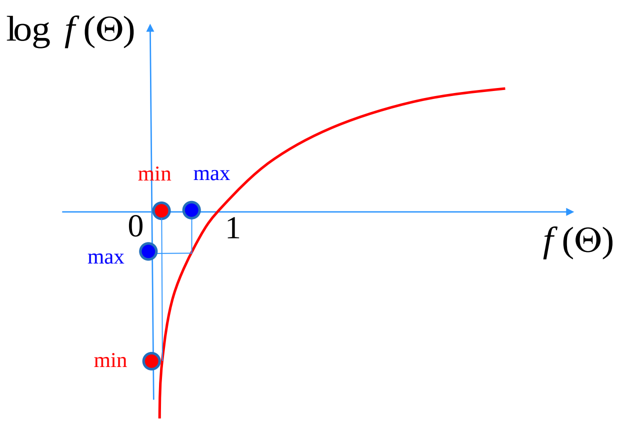
$$\Theta_{ML} = \arg\max_{\Theta} p(D \mid \Theta, \xi)$$

Log-likelihood



Properties of log function: ?

Log-likelihood



$$\Theta^* = \arg\max_{\Theta} f(\Theta) = \arg\max_{\Theta} \log f(\Theta)$$

ML Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta})$ **Data** $D = \{D_1, D_2, .., D_n\}$

$\boxed{\max_{\Theta} p(D \mid \Theta, \xi)}$

- **Maximum likelihood (ML)**
  - Find Θ that maximizes likelihood $p(D \mid \Theta, \xi)$

$$
\begin{aligned}
P(D \mid \Theta, \xi) &= P(D_1, D_2, .., D_n \mid \Theta, \xi) \\
&= P(D_1 \mid \Theta, \xi) P(D_2 \mid \Theta, \xi) \ldots P(D_n \mid \Theta, \xi) \\
&= \prod_{i=1}^{n} P(D_i \mid \Theta, \xi)
\end{aligned}
$$

Independent examples

log-likelihood $\log p(D \mid \Theta, \xi) = \sum_{i=1}^{n} \log P(D_i \mid \Theta, \xi)$

$$\Theta_{ML} = \arg\max_{\Theta} p(D \mid \Theta, \xi) = \arg\max_{\Theta} \log p(D \mid \Theta, \xi)$$

Bayesian parameter estimation

The ML estimate picks just one value of the parameter

- **Problem:** if there are two different parameter values that are close in terms of the likelihood, using only one of them may introduce a strong bias, if we use it, for example, for predictions.

Bayesian parameter estimation

- Remedies the limitation of one choice
- Uses the posterior distribution for parameters Θ
- Posterior 'covers' all possible parameter values (and their "weights")

Parameter posterior Data Likelihood

$$p(\Theta \mid D, \xi) = \frac{p(D \mid \Theta, \xi)\, p(\Theta \mid \xi)}{p(D \mid \xi)}$$

Parameter prior

Bayesian parameter estimation

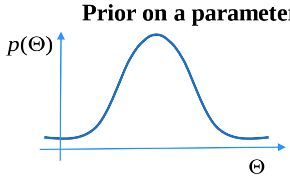
What does it do?

- Prior and Posterior 'covers' all possible parameter values (and their "weights")

Assume: we have a model of $p(x \mid \Theta)$ with a parameter Θ

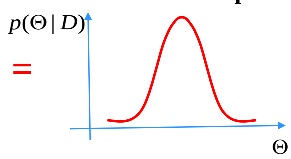
- **Bayesian parameter estimation:**

Prior on a parameter



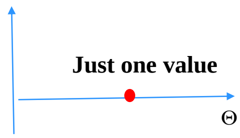
+ **Data +** $p(x \mid \Theta)$ =

Posterior on a parameter



- **ML parameter estimate**

Data + $p(x \mid \Theta)$ =



Bayesian parameter estimation

Bayesian parameter estimation

- Uses the posterior distribution for parameters
- Posterior 'covers' all possible parameter values (and their "weights")

Parameter posterior

Data Likelihood

Parameter prior

$$p(\Theta \mid D, \xi) = \frac{p(D \mid \Theta, \xi)\, p(\Theta \mid \xi)}{p(D \mid \xi)}$$

- **How to use the posterior for modeling** p(X)?

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid D) = \int_{\mathbf{\Theta}} p(X \mid \mathbf{\Theta})\, p(\mathbf{\Theta} \mid D, \xi)\, d\mathbf{\Theta}$$

Parameter estimation

Other criteria:

- **Maximum a posteriori probability (MAP)**

  maximize $p(\mathbf{\Theta} \mid D, \xi)$ (mode of the posterior)

  – Yields: one set of parameters $\mathbf{\Theta}_{MAP}$

  – Approximation:

  $$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta}_{MAP})$$

- **Expected value of the parameter**

  $\hat{\mathbf{\Theta}} = E(\mathbf{\Theta})$ (mean of the posterior)

  – Expectation taken with regard to posterior $p(\mathbf{\Theta} \mid D, \xi)$

  – Yields: one set of parameters

  – Approximation:

  $$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \hat{\mathbf{\Theta}})$$

Parameter estimation. Coin example.

Coin example: we have a coin that can be biased

Outcomes: two possible values -- head or tail

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ

probability of a tail $(1-\theta)$

Objective:

We would like to estimate the probability of a **head** $\hat{\theta}$ from data

Parameter estimation. Example.

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What would be your estimate of the probability of a head ?

$$\tilde{\theta} = ?$$

Parameter estimation. Example

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What would be your choice of the probability of a head ?

Solution: use frequencies of occurrences to do the estimate

$$\tilde{\theta} = \frac{15}{25} = 0.6$$

This is **the maximum likelihood estimate** of the parameter θ

Probability of an outcome

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Assume: we know the probability θ
Probability of an outcome of a coin flip x_i

$$P(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}$$ **Bernoulli distribution**

- Combines the probability of a head and a tail
- So that x_i is going to pick its correct probability
- Gives θ for $x_i = 1$
- Gives $(1-\theta)$ for $x_i = 0$

Probability of a sequence of outcomes.

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Assume: a sequence of independent coin flips

D = H H T H T H **(encoded as D= 110101)**

What is the probability of observing the data sequence **D:**

$$P(D \mid \theta) = ?$$

Probability of a sequence of outcomes.

Data: D a sequence of outcomes x_i such that
- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Assume: a sequence of coin flips D = H H T H T H
encoded as D= 110101

What is the probability of observing a data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

Probability of a sequence of outcomes.

Data: D a sequence of outcomes x_i such that
- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Assume: a sequence of coin flips D = H H T H T H
encoded as D= 110101

What is the probability of observing a data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

likelihood of the data

Probability of a sequence of outcomes

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Assume: a sequence of coin flips D = H H T H T H
encoded as D= 110101

What is the probability of observing a data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

$$P(D \mid \theta) = \prod_{i=1}^{6} \theta^{x_i} (1-\theta)^{(1-x_i)}$$

Can be rewritten using the Bernoulli distribution:

The goodness of fit to the data

Learning: we do not know the value of the parameter θ

Our learning goal:

- Find the parameter θ that fits the data D the best?

Crierion for the best fit: Maximize the likelihood

$$P(D \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)}$$

Intuition:

- more likely are the data given the model, the better is the fit

Note: Instead of an error function that measures how bad the data fit the model we have a measure that tells us how well the data fit :

$$Error(D,\theta) = -P(D \mid \theta)$$

Maximum likelihood (ML) estimate.

Likelihood of data:

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)}$$

Maximum likelihood estimate

$$\Theta_{ML} = \arg\max_{\Theta} p(D \mid \Theta, \xi) = \arg\max_{\Theta} \log p(D \mid \Theta, \xi)$$

Optimize log-likelihood (the same as maximizing likelihood)

$$l(D,\theta) = \log P(D \mid \theta, \xi) = \log \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} =$$

$$\sum_{i=1}^{n} x_i \log\theta + (1-x_i)\log(1-\theta) = \log\theta \underline{\sum_{i=1}^{n} x_i} + \log(1-\theta) \underline{\sum_{i=1}^{n} (1-x_i)}$$

N_1 - number of heads seen $\quad$ N_2 - number of tails seen

Maximum likelihood (ML) estimate.

Optimize log-likelihood

$$l(D,\theta) = N_1 \log\theta + N_2 \log(1-\theta)$$

Set derivative to zero

$$\frac{\partial l(D,\theta)}{\partial\theta} = \frac{N_1}{\theta} - \frac{N_2}{(1-\theta)} = 0$$

Solving $\quad \theta = \dfrac{N_1}{N_1 + N_2}$

ML Solution: $\quad \theta_{ML} = \dfrac{N_1}{N} = \dfrac{N_1}{N_1 + N_2}$

Maximum likelihood estimate. Example

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T

  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of a head and a tail?

Maximum likelihood estimate. Example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T

  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of head and tail ?

Head: $$\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} = \frac{15}{25} = 0.6$$

Tail: $$(1-\theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4$$

Bayesian parameter estimation

Uses the distributions (prior and posterior) over all possible values of the parameter θ of the sampling distribution $p(x \mid \theta)$ (Bernoulli):

Likelihood of data

Posterior

Prior

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)\, p(\theta \mid \xi)}{P(D \mid \xi)}$$

(via Bayes theorem)

Normalizing factor

We know that the likelihood is:

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} = \theta^{N_1} (1-\theta)^{N_2}$$

How to choose the prior probability?

$p(\theta \mid \xi)$ - is the prior probability on θ

Prior distribution

Choice of prior: Beta distribution

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1}$$

$\Gamma(x)$ - a Gamma function $\Gamma(x) = (x-1)\Gamma(x-1)$

For integer values of x $\Gamma(n) = (n-1)!$

Why to use Beta distribution?

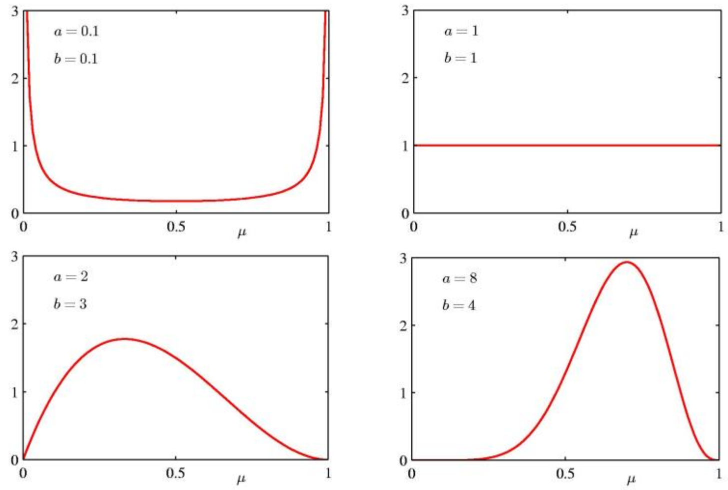
Beta distribution "**fits**" Bernoulli sample - **conjugate choices**

$$P(D \mid \theta, \xi) = \theta^{N_1} (1-\theta)^{N_2}$$

Posterior distribution is again a Beta distribution

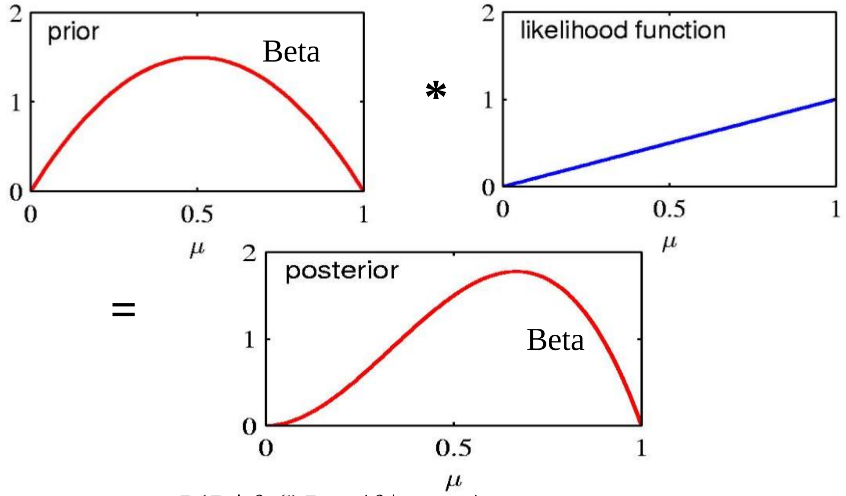
$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

Beta distribution



$$p(\theta \mid \xi) = Beta(\theta \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

Posterior distribution



$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi) Beta(\theta \mid \alpha_1,\alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

Posterior distribution

Beta posterior

- **A conjugate prior to Bernoulli sample**

$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi)Beta(\theta \mid \alpha_1,\alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_2 + N_2)} \theta^{N_1+\alpha_1-1}(1-\theta)^{N_2+\alpha_2-1}$$

Notice that parameters of the prior act like counts of heads and tails (sometimes they are also referred to as **prior counts**)

Maximum aposterior probability (MAP)

Maximum a posteriori estimate

- Selects **the mode of the posterior distribution**

$$\boxed{\theta_{MAP} = \arg\max_{\theta} \, p(\theta \mid D,\xi)}$$

Likelihood of data

prior

$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi)\,p(\theta \mid \xi)}{P(D \mid \xi)}$$

(via Bayes rule)

Normalizing factor

- **The model of the posterior is represented by a Beta distribution**

$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi)Beta(\theta \mid \alpha_1,\alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

Maximum posterior probability

Maximum a posteriori estimate

- Selects the mode of the **posterior distribution**
- **Assumes conjugate prior to Bernoulli sample**

$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi) Beta(\theta \mid \alpha_1,\alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_2 + N_2)} \theta^{N_1+\alpha_1-1}(1-\theta)^{N_2+\alpha_2-1}$$

Mode of the posterior satisfies : $\frac{\partial \log\ p(\theta \mid D,\xi)}{\partial \theta} = 0$

MAP Solution: $\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$

MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate?

MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**
  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate ?

$$\theta_{MAP} = \frac{N_1 + \alpha_1 - 1}{N - 2} = \frac{N_1 + \alpha_1 - 1}{N_1 + N_2 + \alpha_1 + \alpha_2 - 2} = \frac{19}{33}$$

MAP estimate example

- Note that the prior and data fit (data likelihood) are combined
- **The MAP can be biased with large prior counts**
- **It is hard to overturn it with a smaller sample size**
- **Data:**
  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume

$$p(\theta \mid \xi) = Beta(\theta \mid 5,5) \qquad \theta_{MAP} = \frac{19}{33}$$

$$p(\theta \mid \xi) = Beta(\theta \mid 5,20) \qquad \theta_{MAP} = \frac{19}{48}$$

Bayesian framework

- **Predictive probability of an outcome $x=1$ in the next trial**
  $P(x=1|D,\xi)$

$$P(x=1|D,\xi)=\int_0^1 P(x=1|\theta,\xi)\overbrace{p(\theta|D,\xi)}^{\text{Posterior density}}d\theta$$

$$=\int_0^1 \theta p(\theta|D,\xi)d\theta = E(\theta)$$

- **Equivalent to the expected value of the parameter**
  - expectation is taken with respect to the posterior distribution

$$p(\theta|D,\xi)=Beta(\theta|\alpha_1+N_1,\alpha_2+N_2)$$

Expected value of the parameter

How to calculate the expected value of Beta?

$$E(\theta)=\int_0^1 \theta Beta(\theta|\eta_1,\eta_2)d\theta=\int_0^1 \theta \frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)}\theta^{\eta_1-1}(1-\theta)^{\eta_2-1}d\theta$$

$$=\frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)}\int_0^1 \theta^{\eta_1}(1-\theta)^{\eta_2-1}d\theta$$

$$=\frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)}\frac{\Gamma(\eta_1+1)\Gamma(\eta_2)}{\Gamma(\eta_1+\eta_2+1)}\underbrace{\int_0^1 Beta(\eta_1+1,\eta_2)d\theta}_{1}$$

$$=\frac{\eta_1}{\eta_1+\eta_2}$$

Note: $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$ for integer values of α

Expected value of the parameter

- **Substituting the results for the posterior:**

$$p(\theta \mid D, \xi) = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

- **We get** $E(\theta) = \dfrac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$

- **Note that the mean of the posterior is yet** another "reasonable" parameter choice:

$$\hat{\theta} = E(\theta)$$

$$\boxed{\Theta_{EV} = E_{p(\Theta \mid D, \xi)}(\Theta) = \int \Theta p(\Theta \mid D, \xi) d\Theta}$$