CS464
Introduction to Machine Learning

Estimation

(slides based on the slides provided by Öznur Taştan and Mehmet Koyutürk)

Motivation

- In machine learning, we are trying to figure out the relationship between variables (features and outcomes)
  - For this purpose, we use a model (an assumption on the structure of this relationship)
  - A probability distribution usually serves as a good model
  - How do we use observations to learn this distribution?

- Density Estimation
  - Maximum Likelihood Estimator (MLE)
  - Maximum A Posteriori Estimate (MAP)

- Where do we get these probability estimates?

Consider Y=Wealth, X=<Gender, HoursWorked>

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Density Estimation

- We assume that the variable of interest is sampled from a distribution
- We have some observations on the variable
- How do we use observations to learn the distribution?

Density Estimation

- A billionaire asks you a question:
- **He says:** I have a thumbtack, if I flip it, what's the probability it will fall with the nail up (heads)?

- **You say:** Please flip it a few times…

Data

- The billionaire flips the thumbnail 5 times:

- **You say** the probability that it falls with the nail up

$$P(Heads) = 3/5$$

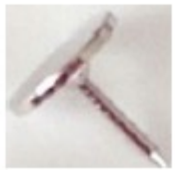
- Why frequency of heads?
- How good is this estimation?
- Why is this a machine learning problem?

Why frequency of heads?

- Frequency of heads is exactly the ***maximum likelihood estimator*** for this problem

Thumbtack- Bernoulli Trial

$P(Heads) = \theta$ and $P(Tails) = 1 - \theta$

D={     ...  }

Thumbtack- Bernoulli Trial

$P(Heads) = \theta$ and $P(Tails) = 1 - \theta$

D={  }

- Flips produce a data set *D*

- Flips are independent, identically distributed and each is a *Bernoulli trial.*

Thumbtack- Bernoulli Trial

$$P(Heads) = \theta \text{ and } P(Tails) = 1 - \theta$$

D={  }

- Flips produce a data set *D*

- Flips are independent, identically distributed and each is a *Bernoulli trial.*

- **Maximum Likelihood Estimator (MLE):** Choose θ that maximizes the probability of observed data

Estimation vs. Learning

- Density estimation is a learning problem too:
  - **Data:** Observed set of flips with with α_H heads and α_T tails
  - **Model:** Bernoulli distribution
  - **Learning:** Finding θ , which is an optimization problem

- Once we estimate θ, we can predict the probability of the next flip being a head
  – We can do more than that too: For example, predict the number of heads in the next 100 flips

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data (likelihood of the data)

The likelihood of observing this data is the joint probability:

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximum likelihood estimate of θ :

$$\boxed{\widehat{\theta} \;=\; \arg\max_{\theta} \; P(\mathcal{D} \mid \theta)}$$

Your First Parameter Learning Algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Why do we take the log?

Your First Parameter Learning Algorithm

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \ \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Why do we take the log?
  - Joint probabilities are often in the form of multiplications and exponents (comes from the independence assumption)
  - Log transforms multiplication to addition
  - The resulting equations are easier to manage

- Take derivative and set it to 0

Your First Parameter Learning Algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Take derivative and set it to 0

$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = \frac{d}{d\theta} \left[\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\right]$$

Your First Parameter Learning Algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \ \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Take derivative and set it to 0

$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = \frac{d}{d\theta}\left[\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\right]$$

$$= \frac{d}{d\theta}\left[\alpha_H \ln \theta + \alpha_T \ln(1-\theta)\right]$$

Your First Parameter Learning Algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln \theta^{\alpha_H} (1-\theta)^{\alpha_T}\end{aligned}$$

- Take derivative and set it to 0

$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = \frac{d}{d\theta} [\ln \theta^{\alpha_H} (1-\theta)^{\alpha_T}]$$

$$= \frac{d}{d\theta} [\alpha_H \ln \theta + \alpha_T \ln(1-\theta)]$$

$$= \alpha_H \frac{d}{d\theta} \ln \theta + \alpha_T \frac{d}{d\theta} \ln(1-\theta)$$

Your First Parameter Learning Algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \ \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Take derivative and set it to 0

$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = \frac{d}{d\theta}\left[\ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\right]$$

$$= \frac{d}{d\theta}\left[\alpha_H \ln \theta + \alpha_T \ln(1-\theta)\right]$$

$$= \alpha_H \frac{d}{d\theta} \ln \theta + \alpha_T \frac{d}{d\theta} \ln(1-\theta)$$

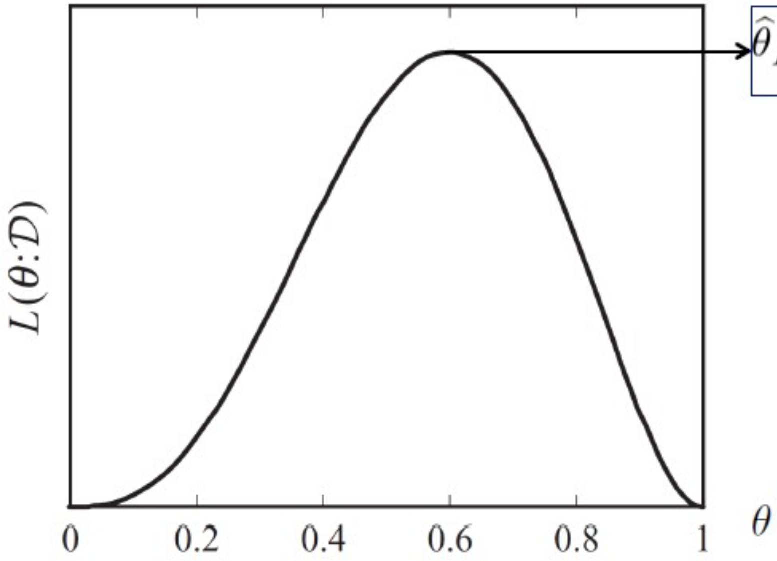
$$= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0$$

$$\boxed{\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}}$$

Maximum Likelihood Estimation

Data 

$$L(\theta; \mathcal{D}) = \ln P(\mathcal{D}|\theta)$$



$$\boxed{\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}}$$

How Many Flips Do I Need?

- Your answer to the billionaire

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- **He says: "**While you have been calculating, I flipped 50 times, 30 times it was head". He asks what is your answer now?

How Many Flips Do I Need?

- Your answer to the billionaire

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- **He says: “**While you have been calculating, I flipped 50 times, 30 times it was head”. He asks what is your answer now?
- **You say:** 30 / 50 = 3/5
- **He says:** Did I waste my time flipping more?

How Many Flips Do I Need?

- Your answer to the billionaire

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- **He says:** **“**While you have been calculating, I flipped 50 times, 30 times it was head”. He asks what is your answer now?
- **You say:** 30 / 50 = 3/5
- **He says:** Did I waste my time flipping more?
- **You say:** No! On the contrary, the more data the merrier
- This is why….

A Bound (from Hoeffding's Inequality)

- Let $N = \alpha_H + \alpha_T$, and $\widehat{\theta}_{MLE} = \dfrac{\alpha_H}{\alpha_H + \alpha_T}$
- Let θ^* be the true parameter. For any $\varepsilon > 0$,

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \;\leq\; 2e^{-2N\epsilon^2}$$

Probably Approximately Correct

PAC: Probably Approximate Correct

Billionaire says: I want to know the thumbtack θ, within ε = 0.1, with probability at least 1-δ = 0.95.

How many flips? Or, how big do I set N?

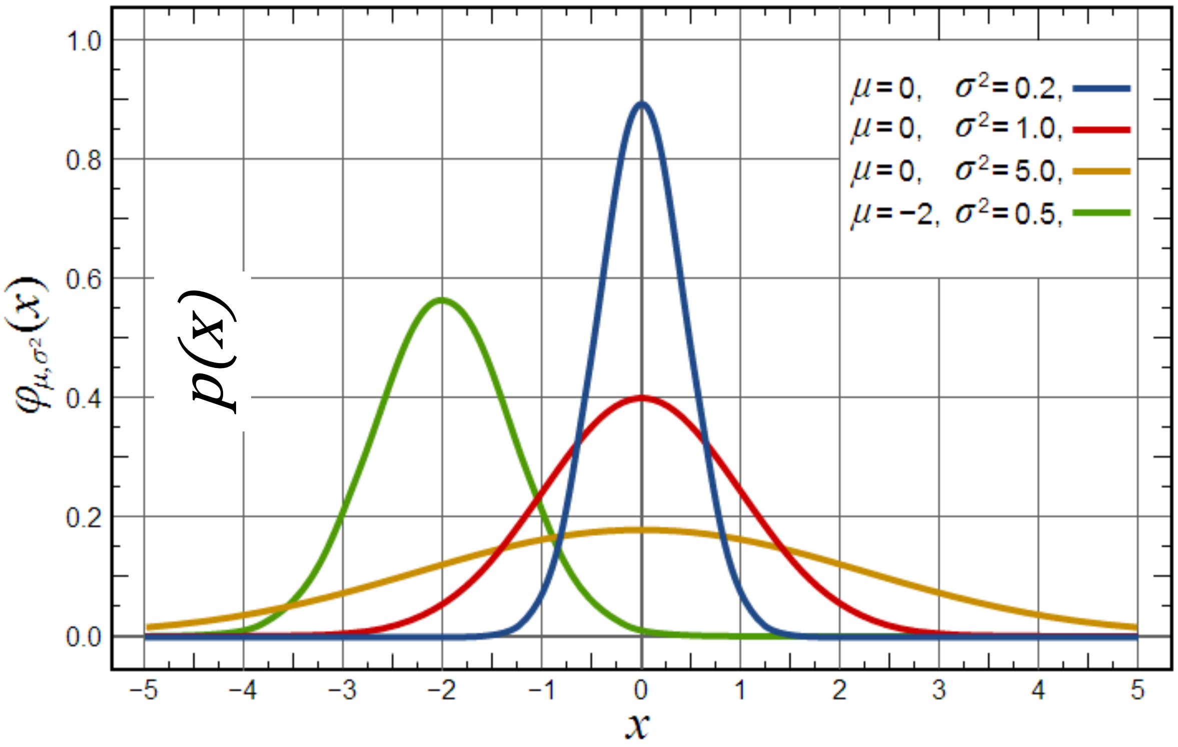
$$P(|\ \widehat{\theta} - \theta^* \ | \geq \underbrace{\epsilon}_{.1}) \ \leq \ \underbrace{2e^{-2N\epsilon^2}}_{= .05}$$

$$N \geq \frac{\ln(2/0.05)}{2 \times 0.1^2} \approx \frac{3.8}{0.02} = 190$$

What if we have a continuous variable?

- What if we are measuring a continuous variable?

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



Learning Parameters For a Gaussian

- Assume we have i.i.d data
- Learn the parameters
  - The mean, μ
  - Standard deviation, σ

x_i	Exam Scores
0	80
1	70
2	12
…	
3	99

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Learning a Gaussian Distribution

- Prob. of i.i.d. samples *D*={x_1,…,x_N}:

$$P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\mu_{MLE}, \sigma_{MLE} = \arg\max_{\mu,\sigma} P(\mathcal{D} \mid \mu, \sigma)$$

Learning a Gaussian Distribution

- Prob. of i.i.d. samples D={x_1,…,x_N}:

$$P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\mu_{MLE}, \sigma_{MLE} = \arg\max_{\mu,\sigma} P(\mathcal{D} \mid \mu, \sigma)$$

- Log-likelihood of data:

$$\begin{aligned} \ln P(\mathcal{D} \mid \mu, \sigma) &= \ln \left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right] \\ &= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

MLE for the Mean

$$\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

MLE for the Mean

$$
\begin{aligned}
\frac{d}{d\mu}\ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\mu}\left[-N\ln\sigma\sqrt{2\pi} - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&= \frac{d}{d\mu}\left[-N\ln\sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N}\frac{d}{d\mu}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]
\end{aligned}
$$

MLE for the Mean

$$\begin{aligned}
\frac{d}{d\mu}\ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&= \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N} \frac{d}{d\mu}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&= \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = 0
\end{aligned}$$

MLE for the Mean

$$
\begin{aligned}
\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&= \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N} \frac{d}{d\mu}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&= \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = 0 \\
&= \sum_{i=1}^{N} x_i - N\mu = 0
\end{aligned}
$$

MLE for the Mean

$$
\begin{aligned}
\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}\right] \\
&= \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N} \frac{d}{d\mu}\left[\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\
&= \sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^2} = 0 \\
&= \sum_{i=1}^{N} x_i - N\mu = 0
\end{aligned}
$$

$$
\boxed{\widehat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i}
$$

MLE for the Variance

$$\frac{d}{d\sigma} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

MLE for the Variance

$$
\begin{aligned}
\frac{d}{d\sigma} \ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&= \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N} \frac{d}{d\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]
\end{aligned}
$$

MLE for the Variance

$$
\begin{aligned}
\frac{d}{d\sigma} \ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}\right] \\
&= \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N} \frac{d}{d\sigma}\left[\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\
&= -\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^3} \qquad = 0
\end{aligned}
$$

MLE for the Variance

$$\begin{aligned}\frac{d}{d\sigma}\ln P(\mathcal{D}\mid\mu,\sigma) &= \frac{d}{d\sigma}\left[-N\ln\sigma\sqrt{2\pi}-\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right]\\ &= \frac{d}{d\sigma}\left[-N\ln\sigma\sqrt{2\pi}\right]-\sum_{i=1}^{N}\frac{d}{d\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]\\ &= -\frac{N}{\sigma}+\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^3} \quad = 0\end{aligned}$$

$$\boxed{\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\widehat{\mu})^2}$$

MLE of Gaussian Parameters

$$\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

The MLE for the variance of a Gaussian is biased. That is, the expected value of the estimator is not equal to the true parameter. An unbiased variance estimator:

$$\widehat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

What if we have prior beliefs?

- **Billionaire** says *wait, I think the thumbtack is close to 50-50*. How can you use this information?

- **You say:** I can learn it the Bayesian way.

Bayesian Rule

- What if we have prior beliefs?

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \frac{\overbrace{p(D|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(D)}_{\text{normalization}}}$$

Utilizing prior information

- θ = probability of landing of the thumptack on heads.
- $\alpha_T = 2$ and $\alpha_H = 1$

Utilizing prior information

- θ = probability of landing of the thumptack on heads.
- $\alpha_T = 2$ and $\alpha_H = 1$
- You have prior knowledge that θ can only take two values: 0.3 or 0.6.

Utilizing prior information

- θ = probability of landing of the thumptack on heads.
- $\alpha_T = 2$ and $\alpha_H = 1$
- You have prior knowledge that θ can only take two values: 0.3 or $0.6.$
- Additionally, you have the prior beliefs on θ :
  $\mathbf{P}\left(\theta = 0.3\right) = 0.2$ and $\mathbf{P}\left(\theta = 0.6\right) = 0.8$.
- How would you take the priors into account?

Bayesian Rule

likelihood

prior

posterior

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

normalization

Maximum A Posteriori (MAP) Estimation

- θ = probability of landing of the thumptack on heads.
- $\alpha_T = 2$ and $\alpha_H = 1$
- You have prior knowledge that θ can only take two values: 0.3 or 0.6
- Additionally, you have the prior beliefs on θ :
  $\mathbf{P}(\theta = 0.3) = 0.2$ and $\mathbf{P}(\theta = 0.6) = 0.8$.
- How would you take the priors into account?

$$\hat{\theta}_{MLE} = \underset{\theta \in \{0.3, 0.6\}}{\arg\max}\ \mathbf{P}(\mathcal{D} \mid \theta)$$

$$\hat{\theta}_{MAP} = \underset{\theta \in \{0.3, 0.6\}}{\arg\max}\ \frac{\mathbf{P}(D \mid \theta)\,\mathbf{P}(\theta)}{\mathbf{P}(D)}$$

Maximum A Posteriori(MAP) Approximation

- θ = probability of landing of the thumptack on heads.
- $\alpha_T = 2$ and $\alpha_H = 1$
- You have prior knowledge that θ can only take two values: 0.3 or 0.6.
- Additionally, you have the prior beliefs on θ :
  $\mathbf{P}(\theta = 0.3) = 0.2$ and $\mathbf{P}(\theta = 0.6) = 0.8$.
- How would you take the priors into account?

$$\hat{\theta}_{MAP} = \underset{\theta \in \{0.3, 0.6\}}{\text{arg max}} \frac{\mathbf{P}(D \,|\, \theta)\, \mathbf{P}(\theta)}{\mathbf{P}(D)}$$

$$\mathbf{P}(\theta = 0.3 \,|\, D) \propto (\mathbf{P}(T \,|\, \theta = 0.3))^2 \mathbf{P}(H \,|\, \theta = 0.3)\, \mathbf{P}(\theta = 0.3) = 0.7^2 * 0.3 * 0.2 = 0.0294$$
$$\mathbf{P}(\theta = 0.6 \,|\, D) \propto (\mathbf{P}(T \,|\, \theta = 0.6))^2 \mathbf{P}(H \,|\, \theta = 0.6)\, \mathbf{P}(\theta = 0.6) = 0.4^2 * 0.6 * 0.8 = 0.0768$$

Therefore $\hat{\theta}_{MAP} = 0.6$

MAP estimation

- Our prior could be in the form of a probability distribution

likelihood prior

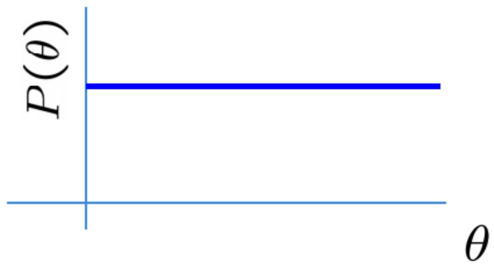
posterior

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

normalization

- Priors can have different forms

- Uninformative prior:
  - Uniform distribution

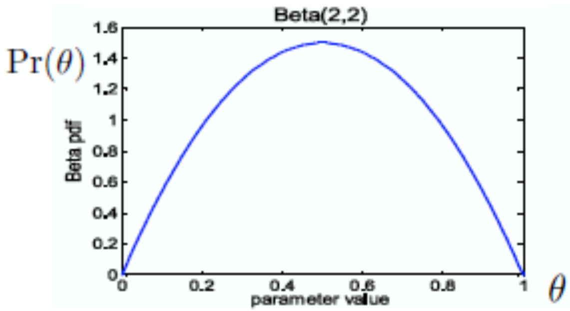


- Conjugate prior:

  - Prior and the posterior have the same form

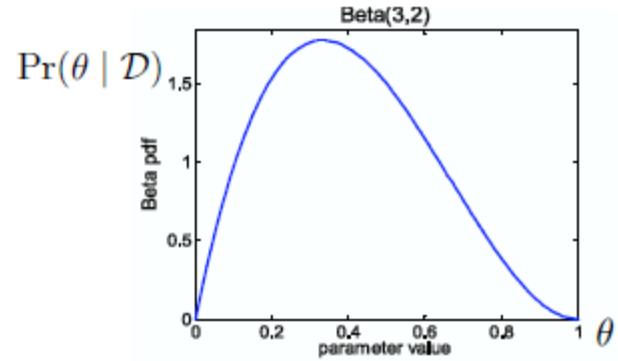
Posterior Distribution

In the beginning



Observe flips
e.g.: {tails, tails}

After observations



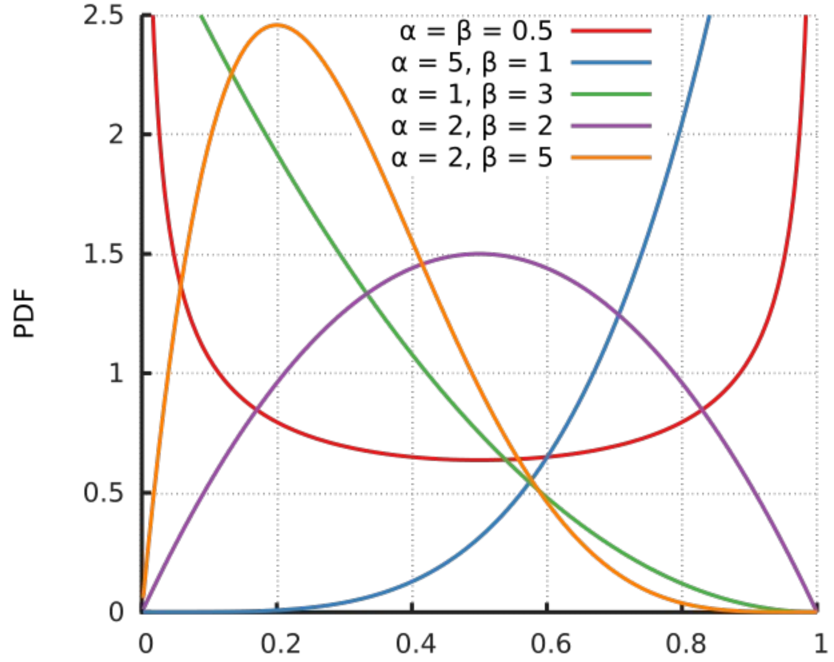
likelihood prior

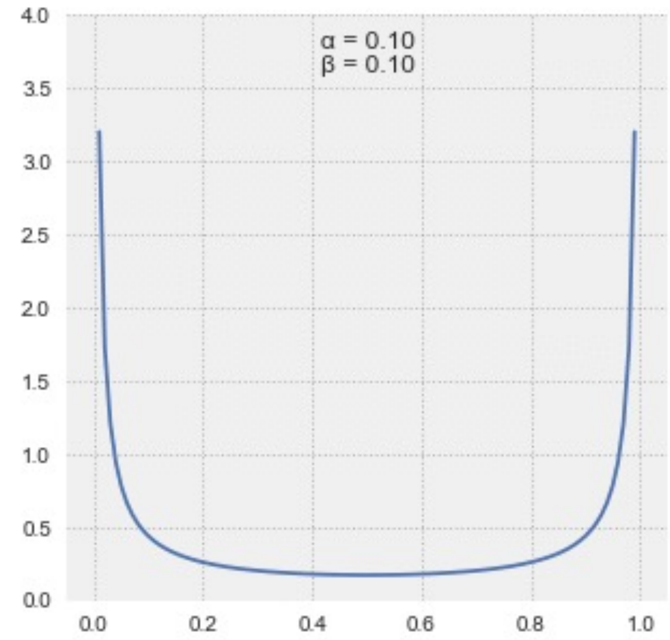
posterior

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

normalization

Beta Distribution





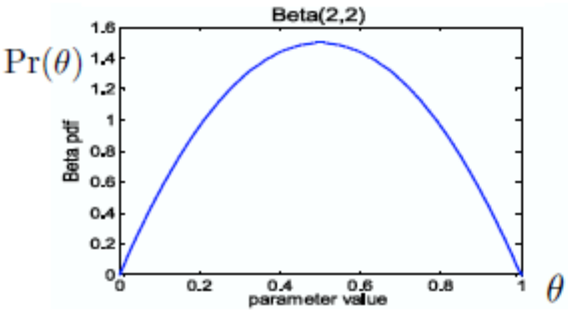
$$0 \leq \theta \leq 1$$

$$\alpha, \beta > 0$$

$$p(\theta) = \frac{1}{B(\alpha, \beta)} * \theta^{\alpha-1} * (1-\theta)^{\beta-1}$$

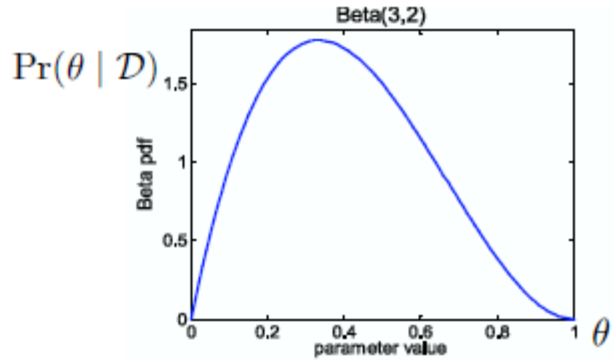
Posterior Distribution

In the beginning



Observe flips
e.g.: {tails, tails}

After observations



likelihood

prior

posterior

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

normalization

$$p(\theta) = \frac{1}{B(\alpha, \beta)} * \theta^{\alpha-1} * (1-\theta)^{\beta-1}$$

Flip it N times, and k times it was head.

N = 3
k = 1

$$p(\theta) = \frac{1}{B(\alpha,\beta)} * \theta^{\alpha-1} * (1-\theta)^{\beta-1}$$

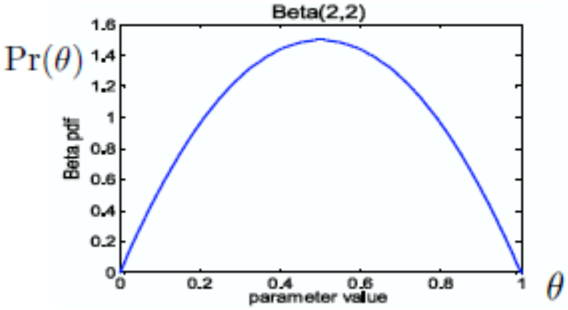
Flip it N times, and k times it was head.

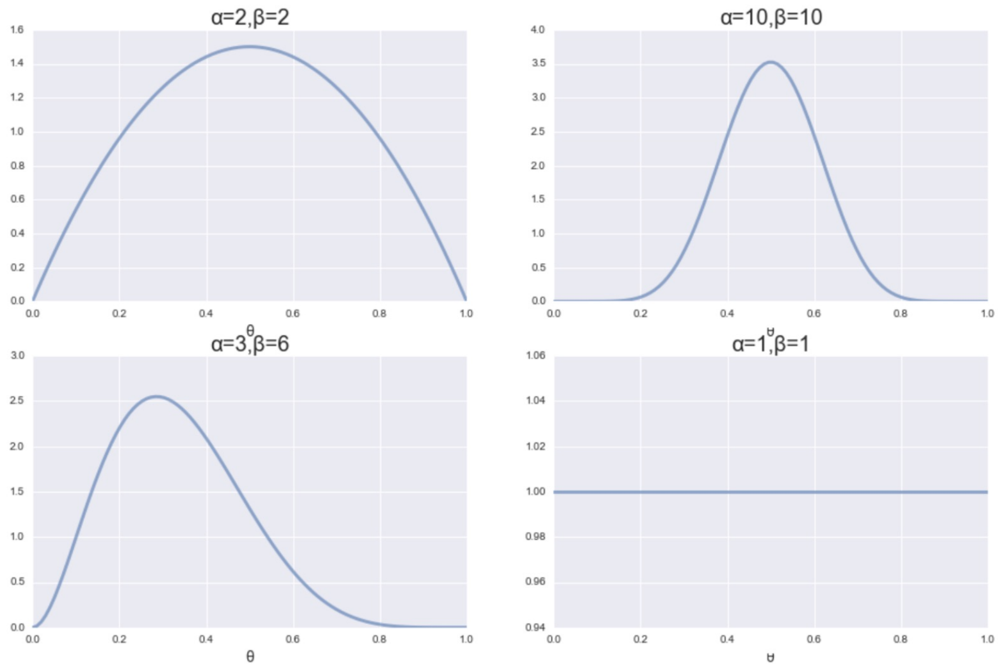
N = 3
k = 1
$\alpha, \beta = 2$

$$\theta_{MLE} = \frac{k}{N} = \frac{1}{3}$$

$$\theta_{MAP} = \frac{k+\alpha-1}{N+\alpha+\beta-2} = \frac{2}{5}$$

In the beginning





$$\theta_{MLE} = \frac{k}{N} = \frac{1}{3}$$

$$\theta_{MAP} = \frac{k + \alpha - 1}{N + \alpha + \beta - 2}$$

Bayesian Estimation

- For the parameters to estimate we assign them an a priori distribution, which is used to capture our prior belief about the parameter
- When the data is sparse, this allows us to fall back to the prior and avoid the issues faced by Maximum Likelihood Estimation (Example: univariate Gaussian)
- When the data is abundant, the likelihood will dominate the prior and the prior will not have much of an effect on the posterior distribution

Estimating Parameters

- Maximum Likelihood Estimate MLE: choose θ that maximizes the probability of observed data $\mathcal{D}$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \mathbf{P}(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given the prior probability of θ and the data $\mathcal{D}$

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} \mathbf{P}(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta} \frac{\mathbf{P}(\mathcal{D} \mid \theta)\,\mathbf{P}(\theta)}{\mathbf{P}(\mathcal{D})}\end{aligned}$$