# Estimating Probabilities from Data

# MLE and MAP

Maria-Florina Balcan

01/29/2018

# Admin

- HWK 1: due on Feb 7th at 5PM.

  Start early!

- Recitation: Jan 30th, 7:00-8:00pm

# What does this has to do with function approximation?

Instead of learning $\mathrm{F}: \mathrm{X} \rightarrow \mathrm{Y}$, learn P(Y|X).

**Can design algorithms that learn functions with uncertain outcomes** (e.g., predicting tomorrow’s stock price) **and that incorporate prior knowledge to guide learning** (e.g., a bias that tomorrow’s stock price is likely to be similar to today’s price).

# The Joint Distribution

- The key to building probabilistic models is to define a set of random variables, and to consider the joint probability distribution over them.

Example: Boolean variables A,B,C

| A | B | C | Prob |
|---|---|---|---|
| 0 | 0 | 0 | 0.30 |
| 0 | 0 | 1 | 0.05 |
| 0 | 1 | 0 | 0.10 |
| 0 | 1 | 1 | 0.05 |
| 1 | 0 | 0 | 0.05 |
| 1 | 0 | 1 | 0.10 |
| 1 | 1 | 0 | 0.25 |
| 1 | 1 | 1 | 0.10 |

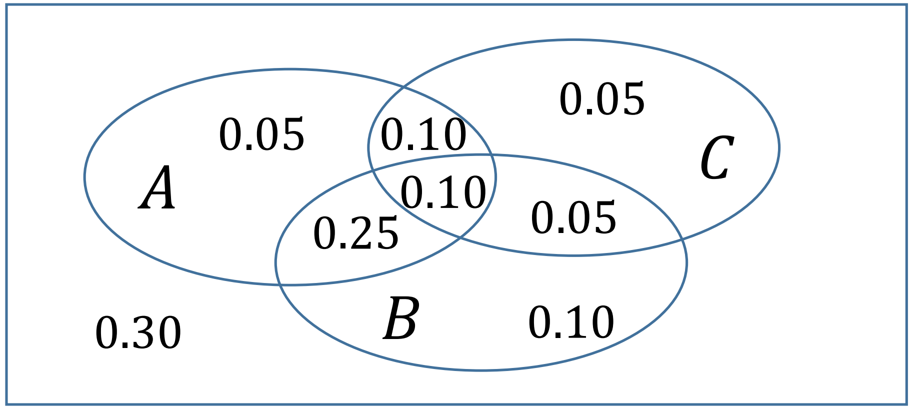
# The Joint Distribution

Recipe for making a joint distribution of $M$ variables:

1. Make a truth table listing all combinations of values ($M$ Boolean variables $\rightarrow 2^M$ rows).
2. For each combination of values, say how probable it is.
3. By the axioms of probability, these probabilities must sum to 1.

Example: Boolean variables A,B,C

| A | B | C | Prob |
|---|---|---|---|
| 0 | 0 | 0 | 0.30 |
| 0 | 0 | 1 | 0.05 |
| 0 | 1 | 0 | 0.10 |
| 0 | 1 | 1 | 0.05 |
| 1 | 0 | 0 | 0.05 |
| 1 | 0 | 1 | 0.10 |
| 1 | 1 | 0 | 0.25 |
| 1 | 1 | 1 | 0.10 |



# Using the Joint Distribution

Once we have the Joint Distribution, can ask for the probability of **any** logical expression involving these variables

| College Degree | Hours worked | Wealth | prob |
|---|---|---|---|
| No | 40.5- | Medium | 0.253122 |
| No | 40.5- | Rich | 0.0245895 |
| No | 40.5+ | Medium | 0.0421768 |
| No | 40.5+ | Rich | 0.0116293 |
| Yes | 40.5- | Medium | 0.331313 |
| Yes | 40.5- | Rich | 0.0971295 |
| Yes | 40.5+ | Medium | 0.134106 |
| Yes | 40.5+ | Rich | 0.105933 |

$$\mathrm{P(E)} = \sum_{\text{rows matching E}} \mathrm{P(row)}$$

# Using the Joint Distribution

Once we have the Joint Distribution, can ask for the probability of **any** logical expression involving these variables

| College Degree | Hours worked | Wealth | prob |
|---|---|---|---|
| No | 40.5- | Medium | 0.253122 |
| No | 40.5- | Rich | 0.0245895 |
| No | 40.5+ | Medium | 0.0421768 |
| No | 40.5+ | Rich | 0.0116293 |
| Yes | 40.5- | Medium | 0.331313 |
| Yes | 40.5- | Rich | 0.0971295 |
| Yes | 40.5+ | Medium | 0.134106 |
| Yes | 40.5+ | Rich | 0.105933 |

$$\mathrm{P(\,College\ \&\ Medium)} = 0.4654$$

$$\mathrm{P(E)} = \sum_{\text{rows matching E}} \mathrm{P(row)}$$

# Using the Joint Distribution

Once we have the Joint Distribution, can ask for the probability of **any** logical expression involving these variables

$$P(\text{Medium}) = 0.7604$$

| College Degree | Hours worked | Wealth | prob |
|---|---|---|---|
| No | 40.5- | Medium | 0.253122 |
| No | 40.5- | Rich | 0.0245895 |
| No | 40.5+ | Medium | 0.0421768 |
| No | 40.5+ | Rich | 0.0116293 |
| Yes | 40.5- | Medium | 0.331313 |
| Yes | 40.5- | Rich | 0.0971295 |
| Yes | 40.5+ | Medium | 0.134106 |
| Yes | 40.5+ | Rich | 0.105933 |

$$P(E) = \sum_{\text{rows matching E}} P(\text{row})$$

# Inference with the Joint Distribution

Once we have the Joint Distribution, can ask for the probability of **any** logical expression involving these variables

| College Degree | Hours worked | Wealth | prob |
|---|---|---|---|
| No | 40.5- | Medium | 0.253122 |
| No | 40.5- | Rich | 0.0245895 |
| No | 40.5+ | Medium | 0.0421768 |
| No | 40.5+ | Rich | 0.0116293 |
| Yes | 40.5- | Medium | 0.331313 |
| Yes | 40.5- | Rich | 0.0971295 |
| Yes | 40.5+ | Medium | 0.134106 |
| Yes | 40.5+ | Rich | 0.105933 |

$$\mathrm{P(College\ |Medium)} = \frac{0.4654}{0.7604} = 0.612$$

$$\mathrm{P(E_1 \mid E_2)} = \frac{\mathrm{P(E_1 \wedge E_2)}}{\mathrm{P(E_2)}} = \frac{\sum_{\text{rows matching } \mathrm{E_1} \text{and } \mathrm{E_2}} \mathrm{P(row)}}{\sum_{\text{rows matching } \mathrm{E_2}} \mathrm{P(row)}}$$

# Learning and the Joint Distribution

Suppose we want to learn the function f: $\langle C, H \rangle \rightarrow W$

Equivalently, $P(W \mid C, H)$

One solution: learn joint distribution from data, calculate $P(W \mid C, H)$

| College Degree | Hours worked | Wealth | prob |
|---|---|---|---|
| No | 40.5- | Medium | 0.253122 |
| No | 40.5- | Rich | 0.0245895 |
| No | 40.5+ | Medium | 0.0421768 |
| No | 40.5+ | Rich | 0.0116293 |
| Yes | 40.5- | Medium | 0.331313 |
| Yes | 40.5- | Rich | 0.0971295 |
| Yes | 40.5+ | Medium | 0.134106 |
| Yes | 40.5+ | Rich | 0.105933 |

e.g., $P(W = \text{rich} | C = \text{no}, H = 40.5 -) = \frac{0.0245895}{0.0245895+0.253122}$

# Idea: learn classifiers by learning $P(Y \mid X)$

Consider $Y = \text{Wealth}$

$X = \langle \text{CollegeDegree}, \text{HoursWorked} \rangle$

| College Degree | Hours worked | Wealth | prob |
|---|---|---|---|
| No | 40.5- | Medium | 0.253122 |
| No | 40.5- | Rich | 0.0245895 |
| No | 40.5+ | Medium | 0.0421768 |
| No | 40.5+ | Rich | 0.0116293 |
| Yes | 40.5- | Medium | 0.331313 |
| Yes | 40.5- | Rich | 0.0971295 |
| Yes | 40.5+ | Medium | 0.134106 |
| Yes | 40.5+ | Rich | 0.105933 |

| College Degree | Hours worked | P(rich\|C,HW) | P(medium\|C,HW) |
|---|---|---|---|
| No | $< 40.5$ | .09 | .91 |
| No | $> 40.5$ | .21 | .79 |
| Yes | $< 40.5$ | .23 | .77 |
| Yes | $> 40.5$ | .38 | .62 |

# Estimating Probabilities from Data

## MLE and MAP

# Estimating the Bias of a Coin

**Problem**: Assume we can flip a coin with bias $\theta$ several times. Estimate the probability that it turns out heads when we flip it?

X=1 X=0

Each flip yields a Boolean value for X, X~ Bernoulli($\theta$)

Bernoulli Random Variable $P(X = 1) = \theta$; $P(X = 0) = 1 - \theta$

We flip it repeatedly, observing the outcome:

- It turns Heads (i.e. X=1) $\alpha_H$ times
- It turns Tails (i.e. X=0) $\alpha_T$ times

How can we estimate the probability of heads $\theta = P(X = 1)$?

# Estimating the Bias of a Coin

**Problem**: Assume we can flip a coin with bias $\theta$ several times. How can we estimate the probability that it turns out heads when we flip it?

We flip it repeatedly, observing the outcome:

- It turns Heads (i.e. X=1) $\alpha_H$ times
- It turns Tails (i.e. X=0) $\alpha_T$ times

How can we estimate the probability of heads $\theta = P(X = 1)$?

Two Cases:

- Case 1: 100 flips. E.g., 51 Heads (X=1) and 49 tails (X=0)
- Case 2: 3 flips. E.g., 2 Heads (X=1) and 1 tails (X=0)

# Principles of Estimating Probabilities

**Principle 1: Maximum Likelihood Estimation** E.g., 51 Heads (X=1) and 49 tails (X=0)

Choose parameter $\hat{\theta}$ that maximizes likelihood of observed data $P(\text{data}|\hat{\theta})$

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

**Principle 2: Maximum Aposteriori Probability** E.g., 2 Heads (X=1) and 1 tails (X=0)

Choose parameter $\hat{\theta}$ that maximizes likelihood the posterior prob $P(\hat{\theta}|\text{data})$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \#\text{halucinated\_Hs}}{(\alpha_T + \#\text{halucinated\_Ts}) + (\alpha_H + \#\text{halucinated\_Hs})}$$

# Principles of Estimating Probabilities

**Principle 1: Maximum Likelihood Estimation**

E.g., 51 Heads (X=1) and 49 tails (X=0)

Choose parameter $\hat{\theta}$ that maximizes likelihood of observed data $P(\text{data}|\hat{\theta})$

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

**Principle 2: Maximum Aposteriori Probability**

E.g., 2 Heads (X=1) and 1 tails (X=0)

Choose parameter $\hat{\theta}$ that maximizes likelihood the posterior prob $P(\hat{\theta}|\text{data})$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \#\text{halucinated\_Hs}}{(\alpha_T + \#\text{halucinated\_Ts}) + (\alpha_H + \#\text{halucinated\_Hs})}$$

# Maximum Likelihood Estimation for Bernoulli Variables

$P(X = 1) = \theta \qquad P(X = 0) = 1 - \theta$

Data D: {1, 0, 0, 1, … }

Flips produce data D with $\alpha_H$ heads (X=1) and $\alpha_T$ tails (X=0)

Flips are i.i.d.:

- independent events
- identically distributed according to the Bernoulli distribution

**MLE estimate: choose the value of θ that makes D most probable.**

Intuition: we are more likely to observe data D if we are in a world where the appearance of this data is highly probable. Therefore, we should estimate θ by assigning it whatever value maximizes the probability of having observed D.

# Maximum Likelihood Estimation for Bernoulli Variables

$P(X = 1) = \theta$ $\qquad$ $P(X = 0) = 1 - \theta$

Data D: {1, 0, 0, 1, … }

Flips produce data D with $\alpha_H$ heads (X=1) and $\alpha_T$ tails (X=0)

Flips are i.i.d.:

- independent events
- identically distributed according to the Bernoulli distribution

Therefore $P(D|\theta) = \theta(1-\theta)(1-\theta)\theta \ldots = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta}\, P(D|\theta)$$

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta}\, \ln P(D|\theta)$$

# Maximum Likelihood Estimation for Bernoulli Variables

$P(X = 1) = \theta \qquad P(X = 0) = 1 - \theta$

Data D: {1, 0, 0, 1, … }  $\alpha_H$ heads and $\alpha_T$ tails

$$\hat{\theta}_{MLE} = \text{argmax}_\theta \ln P(D|\theta)$$
$$= \text{argmax}_\theta \ln[\theta^{\alpha_1}(1-\theta)^{\alpha_0}]$$

Set derivative to 0. $\frac{d}{d\,\theta} \ln P(D|\theta) = 0$

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

$$\frac{d}{d\,\theta} \ln \theta = \frac{1}{\theta}$$

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

# Summary: MLE for Bernoulli Variables

**Problem**: Assume we can flip a coin with bias $\theta$ several times. Estimate the probability that it turns out heads when we flip it?


Each flip yields a Boolean value for X, X~ Bernoulli($\theta$)

Bernoulli Random Variable $\quad \mathrm{P}(\mathrm{X}=1)=\theta; \quad \mathrm{P}(\mathrm{X}=0)=1-\theta$

$$\mathrm{P}(\mathrm{X})=\theta^{\mathrm{X}}(1-\theta)^{1-\mathrm{X}}$$

Data D of independently, identically distributed (i.i.d) flips produces $\alpha_{\mathrm{H}}$ heads (X=1) and $\alpha_{\mathrm{T}}$ tails (X=0)

Therefore $\mathrm{P}(\mathrm{D}|\theta)=(\alpha_1,\alpha_0|\theta)=\theta^{\alpha_{\mathrm{H}}}(1-\theta)^{\alpha_{\mathrm{T}}}$

$$\hat{\theta}_{\mathrm{MLE}}=\operatorname{argmax}_{\theta}\mathrm{P}(\mathrm{D}|\theta)=\frac{\alpha_{\mathrm{H}}}{\alpha_{\mathrm{T}}+\alpha_{\mathrm{H}}}$$

# High Probability Bound, Sample Complexity

**Problem**: Assume we can flip a coin with bias $\theta$ several times. Estimate the probability that it turns out heads when we flip it?

X=1 X=0

$P(X = 1) = \theta$

Data $D: \{1, 0, 0, 1, \dots\}$

$\alpha_H$ heads and $\alpha_T$ tails; $n = \alpha_0 + \alpha_1$

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

**Hoeffding Inequality**:

For any $\epsilon > 0$, $P(|\hat{\theta}_{MLE} - \theta| \geq \epsilon) \leq 2\, e^{-2n\epsilon^2}$

**High Probability Bound**: Want to know the coin parameter $\theta$ within $\epsilon > 0$ with probability at least $1 - \delta$. How many flips?

Set $P(|\hat{\theta}_{MLE} - \theta| \geq \epsilon) \leq 2\, e^{-2n\epsilon^2} \leq \delta$ Solve for n: $n \geq \frac{\ln\frac{2}{\delta}}{2\,\epsilon^2}$

# Principles of Estimating Probabilities

**Principle 1: Maximum Likelihood Estimation**

Choose parameter $\hat{\theta}$ that maximizes likelihood of observed data $P(\text{data}|\hat{\theta})$

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

**Principle 2: Maximum Aposteriori Probability**

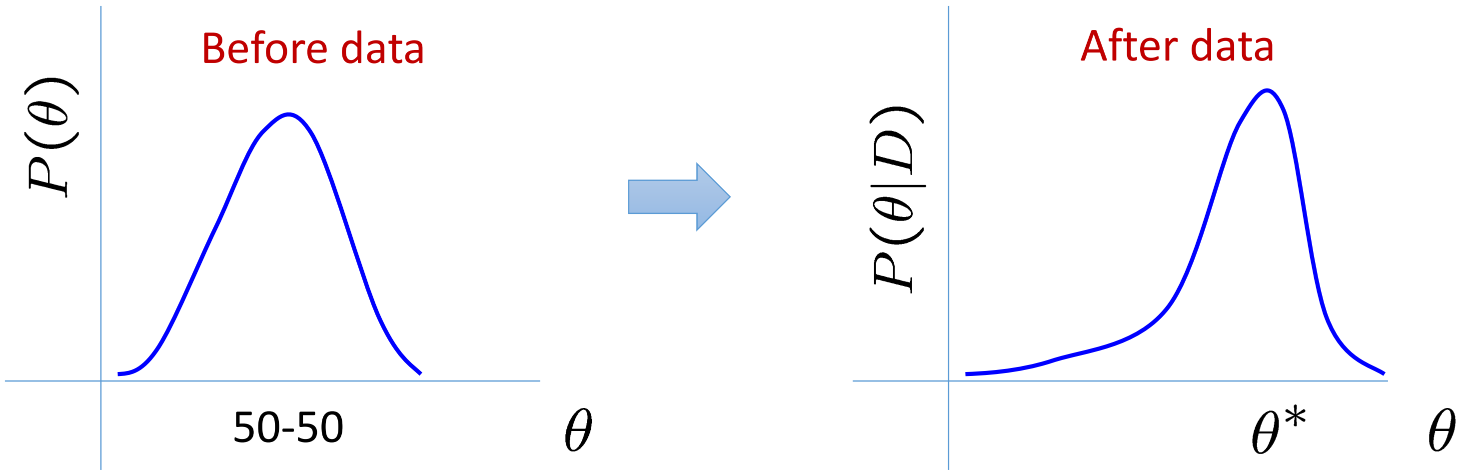
Choose parameter $\hat{\theta}$ that maximizes likelihood the posterior prob $P(\hat{\theta}|\text{data})$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \#\text{halucinated\_Hs}}{(\alpha_T + \#\text{halucinated\_Ts}) + (\alpha_H + \#\text{halucinated\_Hs})}$$

# What if we have prior knowledge?

Prior Knowledge: E.g., I know that the coin is “close” to 50-50.

**MAP estimate: we should choose the value of Theta that is most probable, given the observed data D and our prior assumptions summarized by $\mathbf{P(\theta)}$.**



# Bayesian Learning

Use Bayes Rule:

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

Equivalently:

$$P(\theta|D) \propto P(D|\theta) \cdot P(\theta)$$

posterior likelihood prior

**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

**MAP estimate: choose parameter $\hat{\theta}$ that maximizes the posterior prob $P(\hat{\theta}|\text{data})$, i.e. it chooses the value that is most probable given observed data and prior belief**

# Principles of Estimating Probabilities

**Principle 1: Maximum Likelihood Estimation (MLE)**

Choose parameter $\hat{\theta}$ that maximizes likelihood of observed data $P(D|\hat{\theta})$

$$\hat{\theta}_{MLE} = \mathrm{argmax}_{\theta} P(D|\theta)$$

**Principle 2: Maximum Aposteriori Probability (MAP)**

Choose parameter $\hat{\theta}$ that maximizes likelihood the posterior prob $P(\hat{\theta}|D)$, i.e. it chooses the value that is most probable given observed data and prior belief

$$\hat{\theta}_{MAP} = \mathrm{argmax}_{\theta} P(\theta|D) \quad = \mathrm{argmax}_{\theta}\, P(D|\theta)P(\theta)$$

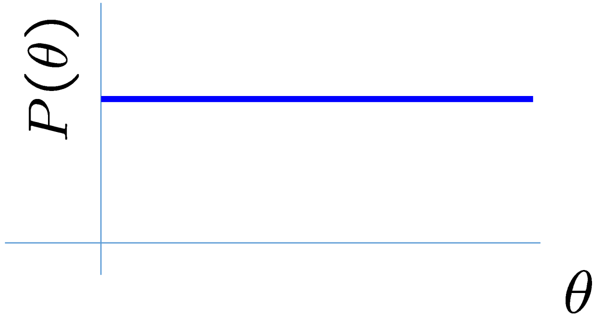
As $n \to \infty$, prior is forgotten

For small sample sizes, prior is important

# Which Prior Distribution?

- Prior represents the experts knowledge.
- Simple posterior form (engineer's approach).

Uninformative Prior

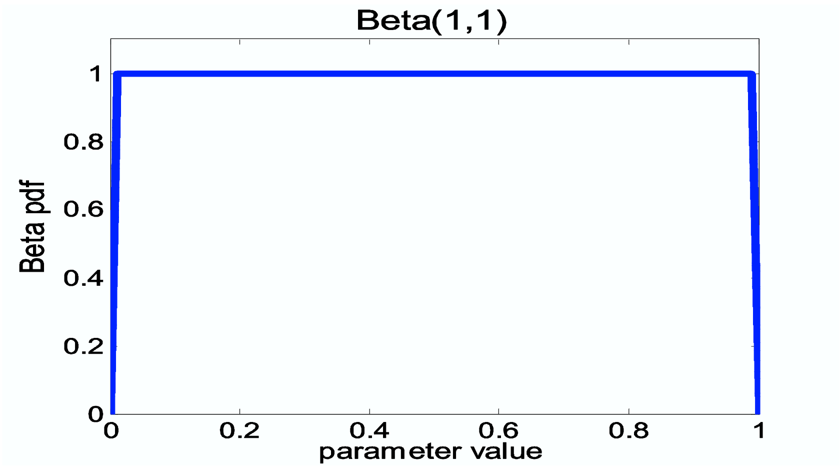


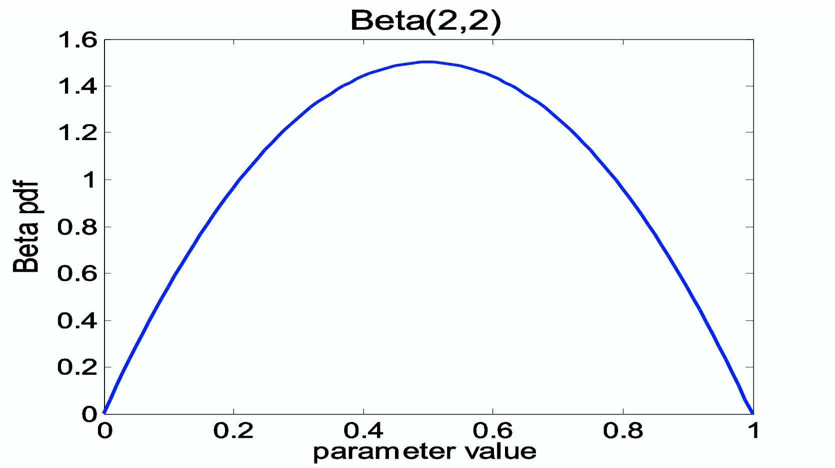

**Conjugate Prior**

- Closed-form expression of posterior.
- $P(\theta)$ and $P(\theta|D)$ have **same** form.

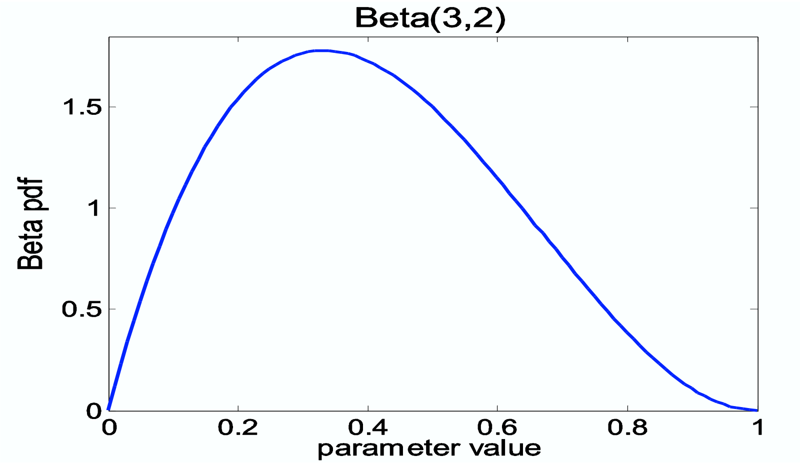
# Beta Prior Distribution

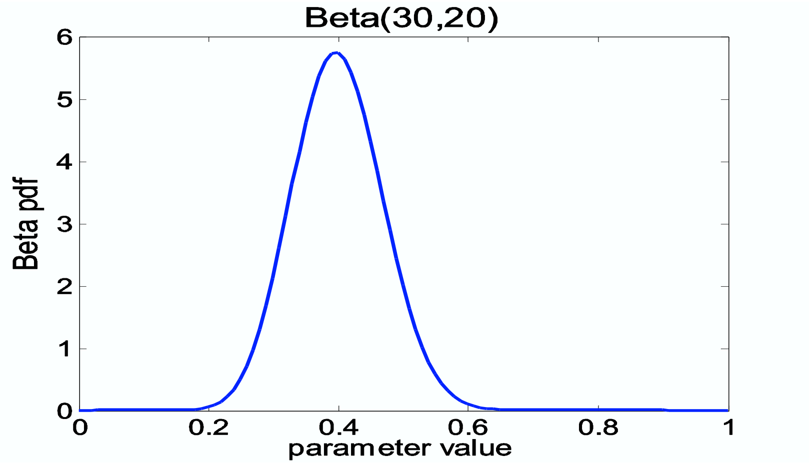
Assume $\theta \sim Beta(\beta_H, \beta_T)$ I.e., $P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)}$




More concentrated as values of $\beta_H$, $\beta_T$ increase



# MAP Estimate for Bernoulli Variables with Beta Prior Distribution


Assume $\theta \sim \mathrm{Beta}(\beta_H, \beta_T)$ I.e., $P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)}$

Likelihood function $P(D|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

Posterior: $P(\theta|D) \propto P(D|\theta)P(\theta)$

$$P(\theta|D) \propto \theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_T - 1}$$

Interpretation: like MLE, but *hallucinating* $\beta_H - 1$ additional heads & $\beta_T - 1$ additional tails

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{(\alpha_T + \beta_T - 1) + (\alpha_H + \beta_H - 1)}$$

Note: as we get more sample effect of prior washed out.

# Conjugate Priors

Likelihood function: $P(D|\theta)$

Prior: $P(\theta)$

Posterior: $P(\theta|D) \propto P(D|\theta)P(\theta)$

Conjugate Prior: $P(\theta)$ is the conjugate prior for the likelihood function $P(D|\theta)$ if the forms of $P(\theta)$ and $P(\theta|D)$ are the same.

# MAP Estimate for Bernoulli Variables with Beta Prior Distribution

Likelihood function $P(D|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$ (Binomial)

If prior is beta distribution, $\theta \sim Beta(\beta_H, \beta_T)$ I.e., $P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)}$

then posterior : $P(\theta|D) \propto P(D|\theta)P(\theta) \propto \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1} \sim Beta(\alpha_H+\beta_H, \alpha_T+\beta_T)$

Therefore

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H}{(\alpha_T+\beta_T - 1) + (\alpha_H+\beta_H - 1)}$$

Mode of Beta distribution

# MAP Estimate for Dice Rolling with Dirichlet Prior Distribution

**Dice Roll Problem: 6 outcomes instead of 2.**

Likelihood function is $\sim \mathrm{Multinomial}(\theta_1, \dots, \theta_k)$ $P(D|\theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2} \cdots \theta_k^{\alpha_k}$

If prior is Dirichlet distribution, $\theta \sim \mathrm{Dirichlet}(\beta_1, \beta_2, \dots, \beta_k)$ $P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \beta_2, \dots, \beta_k)}$

then posterior:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \propto \mathrm{Dirichlet}(\alpha_1 + \beta_1, \dots, \alpha_k + \beta_k)$$

For Multinomial, conjugate prior is Dirichlet.

# Principles of Estimating Probabilities

**Principle 1: Maximum Likelihood Estimation (MLE)**

Choose parameter $\hat{\theta}$ that maximizes likelihood of observed data $P(D|\hat{\theta})$

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta} P(D|\theta)$$

**Principle 2: Maximum Aposteriori Probability (MAP)**

Choose parameter $\hat{\theta}$ that maximizes likelihood the posterior prob $P(\hat{\theta}|D)$, i.e. it chooses the value that is most probable given observed data and prior belief

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} P(\theta|D) \quad = \operatorname{argmax}_{\theta} P(D|\theta)P(\theta)$$

As $n \rightarrow \infty$, prior is forgotten

For small sample sizes, prior is important

# Bayesians vs. Frequentists

You are no good when sample is small

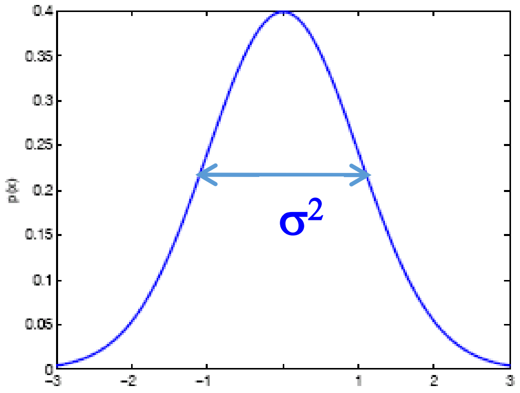
You give a different answer for different priors

# What About Continuous Random Variables?

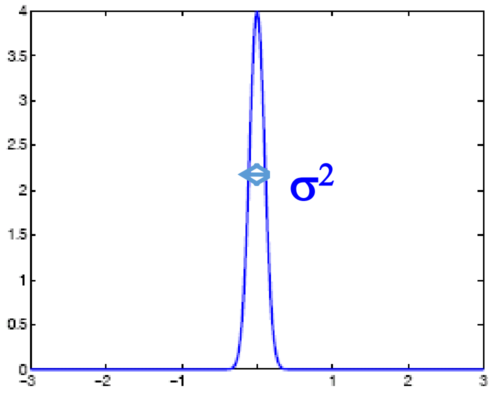
## Gaussian Random Variable

$X \sim N(\mu, \sigma)$, then

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



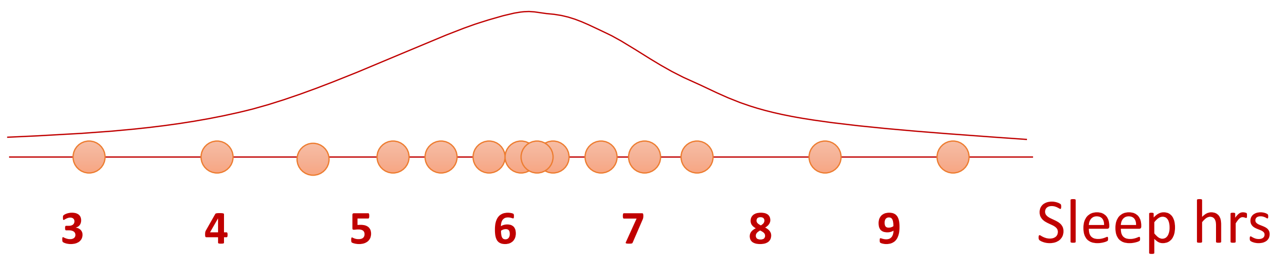

μ=0




μ=0

# What About Continuous Random Variables?

**Observed data D:**




Parameters: $\mu$- mean, $\sigma^2$ variance

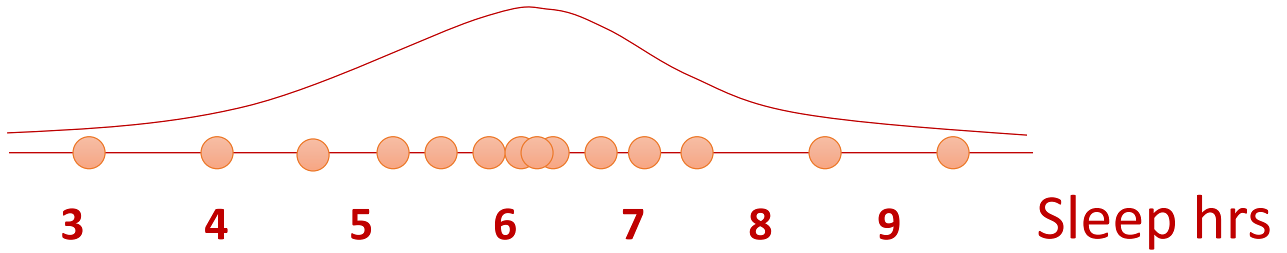
Sleep hours are i.i.d.:

- independent events
- identically distributed according to Gaussian distribution

Goal: estimate $\mu$, $\sigma$

# MLE for Mean of Gaussian

**Observed data D:**




Probability of i.i.d. samples $D = \{x_1, \dots, x_N\}$ $\quad P(D|\mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{\{i=1\dots N\}} e^{-\frac{(x_i-\mu)^2}{\sigma^2}}$

Log-likelihood of data $\quad \ln P(D|\mu,\sigma) = \ln \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{\{i=1\dots N\}} e^{-\frac{(x_i-\mu)^2}{\sigma^2}}$

$$\ln P(D|\mu,\sigma) = -N\ln\left(\sigma\sqrt{2\pi}\right) - \sum_{\{i=1,\dots,N\}} \frac{(x_i-\mu)^2}{\sigma^2}$$

# MLE for Mean of Gaussian

Probability of i.i.d. samples $D = \{x_1, \dots, x_N\}$ $\quad P(D|\mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{\{i=1,\dots,N\}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$\ln P(D|\mu,\sigma) = -N\ln\left(\sigma\sqrt{2\pi}\right) - \sum_{\{i=1,\dots,N\}} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{d}{d\,\mu}\ln P(D|\mu,\sigma) = -\sum_{\{i=1,\dots,N\}} \frac{d}{d\,\mu}\frac{(x_i-\mu)^2}{2\sigma^2} = 2\sum_{\{i=1,\dots,N\}} \frac{(x_i-\mu)}{2\sigma^2}$$

Set $\frac{d}{d\,\mu}\ln P(D|\mu,\sigma) = 0$ $\quad$ Therefore $\sum_{\{i=1,\dots,N\}}(x_i-\mu) = 0$ $\quad$ $\hat{\mu}_{MLE} = \frac{\sum_i x_i}{N}$

# MLE for Variance of Gaussian

Probability of i.i.d. samples $D = \{x_1, \ldots, x_N\}$ $\quad P(D|\mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{\{i=1\ldots N\}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$\ln P(D|\mu,\sigma) = -N\ln\left(\sigma\sqrt{2\pi}\right) - \sum_{\{i=1,\ldots,N\}} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{d}{d\,\sigma}\ln P(D|\mu,\sigma) = -N\frac{d}{d\,\sigma}\ln\left(\sigma\sqrt{2\pi}\right) - \sum_{\{i=1,\ldots,N\}} \frac{d}{d\,\sigma}\frac{(x_i-\mu)^2}{2\sigma^2} = -\frac{N}{\sigma} + 2\sum_{\{i=1,\ldots,N\}} \frac{(x_i-\mu)^2}{2\sigma^3}$$

Set $\frac{d}{d\,\mu}\ln P(D|\mu,\sigma) = 0$ $\qquad$ Therefore $\quad \hat{\sigma}_{MLE} = \frac{\sum_i (x_i - \hat{\mu})^2}{N}$

# Learning Gaussian Parameters

MLE: $$\hat{\sigma}_{\mathrm{MLE}} = \frac{\sum_i (x_i - \mu)^2}{N}$$

$$\hat{\mu}_{\mathrm{MLE}} = \frac{\sum_i x_i}{N}$$

Bayesian learning/estimation is also possible.

Conjugate priors:

Mean: Gaussian prior

Variance: Wishart distribution

# What You Should Know

- MLE, MAP
- Coins, Dice, Gaussian