CSE 446: Point Estimation
Winter 2012

Dan Weld

Some slides from Carlos Guestrin, Luke Zettlemoyer & K Gajos

Logistics

- PS2 out shortly

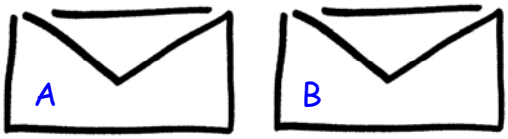
Last Time

- Random variables, distributions
- Marginal, joint & conditional probabilities
- Sum rule, product rule, Bayes rule
- Independence, conditional independence

- Binomial distribution
- Maximum Likelihood Estimation (MLE)
- [A peek at PAC learning theory]

Envelope Problem

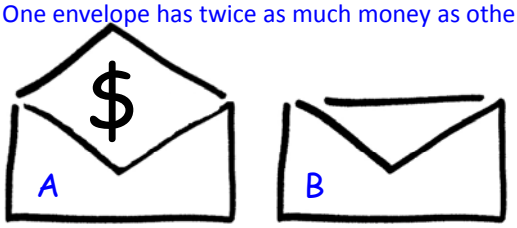
One envelope has twice as much money as other



Switch?

Envelope Problem

One envelope has twice as much money as other



A has $20

E[B] = $ ½ (10 + 40)

= $ 25

Switch?

Coin Flip

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

Which coin will I use?

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Prior: Probability of a hypothesis before we make any observations

Coin Flip



C_1 C_2 C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

Which coin will I use?

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Uniform Prior: All hypothesis are equally likely before we make any observations

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = ?$ $P(C_2|H) = ?$ $P(C_3|H) = ?$

$$P(C_1|H) = \frac{P(H|C_1)P(C_1)}{P(H)} \qquad P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i)$$

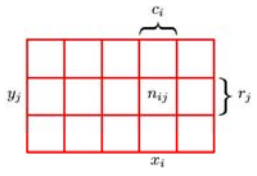


C_1 C_2 C_3

$P(H|C_1)=0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1)=1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

$$P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i)$$



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i)p(X = x_i)$$

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = 0.066$ $P(C_2|H) = 0.333$ $P(C_3|H) = 0.6$

Posterior: Probability of a hypothesis given data



C_1 C_2 C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Terminology

- **Prior**:
  - Probability of a hypothesis before we see any data
- **Uniform Prior**:
  - A prior that makes all hypothesis equally likely
- **Posterior**:
  - Probability of a hypothesis after we saw some data
- **Likelihood**:
  - Probability of data given hypothesis

Experiment 2: Tails

Now, Which coin did I use?

$P(C_1|HT) = ?$ $P(C_2|HT) = ?$ $P(C_3|HT) = ?$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$



C_1 C_2 C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Experiment 2: Tails

Now, Which coin did I use?

$P(C_1|HT) = 0.21$ $P(C_2|HT) = 0.58$ $P(C_3|HT) = 0.21$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

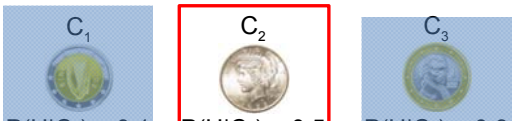
$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = 0.21$ $P(C_2|HT) = 0.58$ $P(C_3|HT) = 0.21$

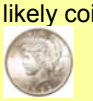


$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin: C_2 

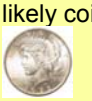
Best estimate for P(H): $P(H|C_2) = 0.5$

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Your Estimate?

Maximum Likelihood Estimate: The best hypothesis that fits observed data assuming uniform prior

Most likely coin: C_2 

Best estimate for P(H): $P(H|C_2) = 0.5$

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

Using Prior Knowledge

- Should we always use a ***Uniform Prior*** ?
- Background knowledge:

  Heads => we have to buy Dan chocolate

  Dan ***likes*** chocolate…

  => Dan is more likely to use a coin biased in his favor

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

Using Prior Knowledge

We can encode it in the prior:

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = ?$ $P(C_2|H) = ?$ $P(C_3|H) = ?$

$$P(C_1|H) = \alpha P(H|C_1)P(C_1)$$


C_1


C_2


C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = 0.006$ $P(C_2|H) = 0.165$ $P(C_3|H) = 0.829$

Compare with ML posterior after Exp 1:

$P(C_1|H) = 0.066$ $P(C_2|H) = 0.333$ $P(C_3|H) = 0.600$


C_1


C_2


C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = ?$ $P(C_2|HT) = ?$ $P(C_3|HT) = ?$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$


C_1


C_2


C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$


C_1


C_2


C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$


C_1


C_2


C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin: C_3 

Best estimate for P(H): $P(H|C_3) = 0.9$


C_1


C_2


C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

Your Estimate?

Maximum A Posteriori (MAP) Estimate:
The best hypothesis that fits observed data assuming a non-uniform prior

Most likely coin: C_3 

Best estimate for P(H)

$P(H|C_3) = 0.9$

C_3 

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

Did We Do The Right Thing?

$P(C_1|HT)=0.035$ $P(C_2|HT)=0.481$ $P(C_3|HT)=0.485$


C_1


C_2


C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

Did We Do The Right Thing?

$P(C_1|HT) =0.035$ $P(C_2|HT)=0.481$ $P(C_3|HT)=0.485$

C_2 and C_3 are almost equally likely


C_1


C_2


C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

A Better Estimate

Recall: $P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i)$ = 0.680

$P(C_1|HT)=0.035$ $P(C_2|HT)=0.481$ $P(C_3|HT)=0.485$


C_1


C_2


C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

Bayesian Estimate

Bayesian Estimate: Minimizes prediction error, given data and (generally) assuming a non-uniform prior

$P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i)$ = 0.680

$P(C_1|HT)=0.035$ $P(C_2|HT)=0.481$ $P(C_3|HT)=0.485$


C_1


C_2


C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

Comparison

After more experiments: HTH^8

ML (Maximum Likelihood):
- P(H) = 0.5
- after 10 experiments: P(H) = 0.9

MAP (Maximum A Posteriori):
- P(H) = 0.9
- after 10 experiments: P(H) = 0.9

Bayesian:
- P(H) = 0.68
- after 10 experiments: P(H) = 0.9

Summary

ML (Maximum Likelihood):

- Easy to compute

MAP (Maximum A Posteriori):

- Still easy to compute
- Incorporates prior knowledge

Bayesian:

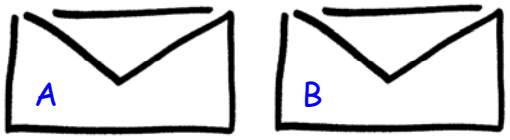
- Minimizes error => great when data is scarce
- Potentially much harder to compute

Recap

	Prior	Hypothesis
Maximum Likelihood Estimate	Uniform	The most likely
Maximum A Posteriori Estimate	Any	The most likely
Bayesian Estimate	Any	Weighted combination

Envelope Problem

One envelope has twice as much money as other

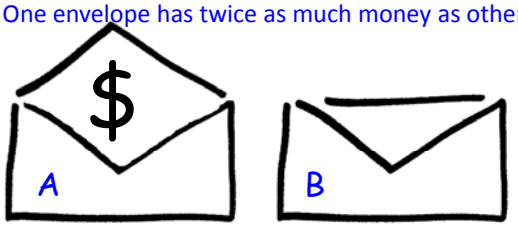


Switch?

33

Envelope Problem

One envelope has twice as much money as other



A has $20

E[B] = $ ½ (10 + 40)

= $ 25

Switch?

34

Those Pesky Distributions

	Discrete		Continuous
	Binary {0, 1}	M Values	
	Bernouilli		Gaussian ~ Normal
	P(x=1) = θ P(x=0) = 1-θ		
Sufficient Statistics	θ		μ, σ

35

Those Pesky Distributions

	Discrete		Continuous
	Binary {0, 1}	M Values	
Single Event	Bernouilli		
Sequence (N trials) N=	Binomial $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$	Multinomial	
Conjugate Prior	Beta	Dirichlet	

36

Those Pesky Distributions

	Discrete		Continuous
	Binary {0, 1}	M Values	
Single Event	Bernouilli		
Sequence (N trials)	Binomial $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$	Multinomial	
N= $\alpha_H+\alpha_T$	↓	↓	
Conjugate Prior	Beta	Dirichlet	

37

Prior Distributions

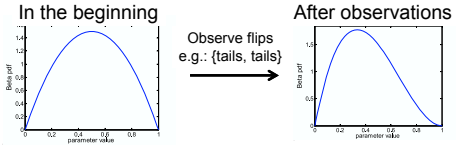
Slightly harder…

- ***Which coin*** is he using (of three known choices)?
- What is the ***bias of a single new, unseen coin***?

38

What if I have prior beliefs?

- Billionaire says: Here's a new coin; I bet it's "close" to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**
- Rather than estimating a single θ, we obtain a distribution over possible values of θ



Bayesian Learning

Use Bayes rule:

Data Likelihood

Prior

Posterior

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

Normalization

Or equivalently: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$

Remember, for uniform priors:

→ reduces to MLE objective

$P(\theta) \propto 1$ $\quad P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)$

Bayesian Learning for Dollars

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

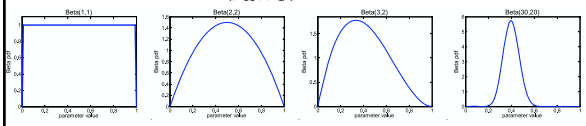
Likelihood function is Binomial:

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

- What about prior?
  - Represent expert knowledge
  - ***Want*** simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior
  - **For Binomial, conjugate prior is Beta distribution**

Beta prior distribution: P(θ)

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



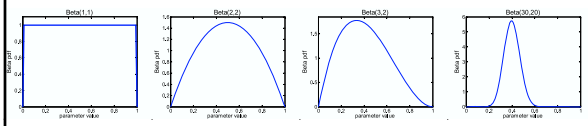
- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$

$$P(\theta \mid \mathcal{D}) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}\ \theta^{\beta_H-1}(1-\theta)^{\beta_T-1}$$
$$= \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_t+1}$$
$$= Beta(\alpha_H+\beta_H, \alpha_T+\beta_T)$$

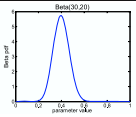
Posterior Distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails
- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



Bayesian Posterior Inference



- Posterior distribution:

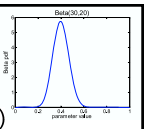
$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- Bayesian inference:
  - No longer single parameter
  - For any specific f, the function of interest
  - Compute the expected value of f

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

  - Integral is often hard to compute

MAP: Maximum a Posteriori Approximation



$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

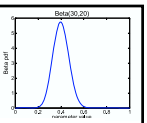
$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

- As more data is observed, Beta is more certain
- MAP: use most likely parameter to approximate the expectation

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

$$E[f(\theta)] \approx f(\hat{\theta})$$

MAP for Beta distribution



$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1}(1-\theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

MAP: use most likely parameter:

$$\hat{\theta} = \arg\max_\theta P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

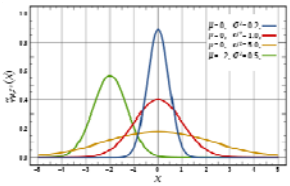
Beta prior equivalent to extra thumbtack flips

As $N \to \infty$, prior is "forgotten"

But, for small sample size, prior is important!

What about continuous variables?

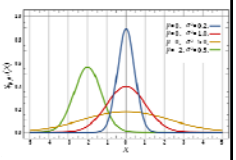
- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**



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

Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant) are Gaussian
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \rightarrow Y \sim N(a\mu+b,a^2\sigma^2)$
- Sum of Gaussians is Gaussian
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$
- Easy to differentiate, as we will see soon!



Learning a Gaussian

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores
- Learn parameters
  - Mean: μ
  - Variance: σ

X_i $i =$	Exam Score
0	85
1	95
2	100
3	12
…	…
99	89

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

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

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_N\}$:

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

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

- Log-likelihood of data:

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

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

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

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

MLE for variance

- Again, set derivative to zero:

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

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

Learning Gaussian parameters

- MLE:

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

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

- BTW. MLE for the variance of a Gaussian is **biased**
  - Expected result of estimation is **not** true parameter!
  - Unbiased variance estimator:

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

Bayesian learning of Gaussian parameters

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution
- Prior for mean:

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

MAP for mean of Gaussian

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

$$\frac{d}{d\mu}[\ln P(\mathcal{D} \mid \mu) P(\mu)] = \frac{d}{d\mu}[\ln P(\mathcal{D} \mid \mu) + \ln P(\mu)]$$

$$\frac{\partial}{\partial \mu} \ln P(\mathcal{D}|\mu) = \sum_i \frac{(x_i - \mu)}{\sigma^2}$$

$$\frac{\partial}{\partial \mu} \ln\left[\frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{(\mu-\eta)^2}{2\lambda^2}}\right) = -\frac{(\mu-\eta)}{\lambda^2}$$

$$\sum_i \frac{(x_i - \mu)}{\sigma^2} + \frac{(\eta - \mu)}{\lambda^2} = 0$$

$$\frac{N\mu}{\sigma^2} + \frac{\mu}{\lambda^2} = \sum_i \frac{x_i}{\sigma^2} + \frac{\eta}{\lambda^2}$$

$$\Rightarrow \mu = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\lambda^2}} \left[\sum_i \frac{x_i}{\sigma^2} + \frac{\eta}{\lambda^2}\right]$$