Learning Distributions (Parametric Approach)

Aarti Singh

Machine Learning 10-701/15-781
Sept 12, 2012

MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Announcement

- HW 1 is out – Due Sept 26

Your first consulting job

- A billionaire asks you a question:
  - He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
  - You say: Please flip it a few times:

  - You say: The probability is:
  - **He says: Why???** **3/5**
  - You say: Because…

Bernoulli distribution

Data, *D* =

$$D = \{X_i\}_{i=1}^n,\ X_i \in \{\mathrm{H}, \mathrm{T}\}$$

- P(Heads) = θ, P(Tails) = 1-θ
- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

<u>Choose θ that maximizes the probability of observed data</u>

Maximum Likelihood Estimation

Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

MLE of probability of head:

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$ = 3/5 "Frequency of heads"

Number of heads   Number of tails

Maximum Likelihood Estimation

Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta)$$

Identically distributed

$$= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

Maximum Likelihood Estimation

Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

$$= \arg\max_{\theta} \; \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

$$\frac{\partial J(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\alpha_H(1-\theta) - \alpha_T\theta \Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

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

How many flips do I need?

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

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 3/5, I can prove it!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **He says: What's better?**
- You say: Hmm… The more the merrier???
- He says: Is this why I am paying you the big bucks???

Simple bound (Hoeffding's inequality)

- For $n = \alpha_H + \alpha_T$, and $\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

- Let θ^* be the true parameter, for any $\varepsilon > 0$:

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

PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the coin parameter θ, within $\varepsilon = 0.1$, with probability at least $1\text{-}\delta = 0.95$. How many flips?

$$P(|\ \widehat{\theta} - \theta^* \ |\geq \epsilon) \quad \leq \quad 2e^{-2n\epsilon^2}$$

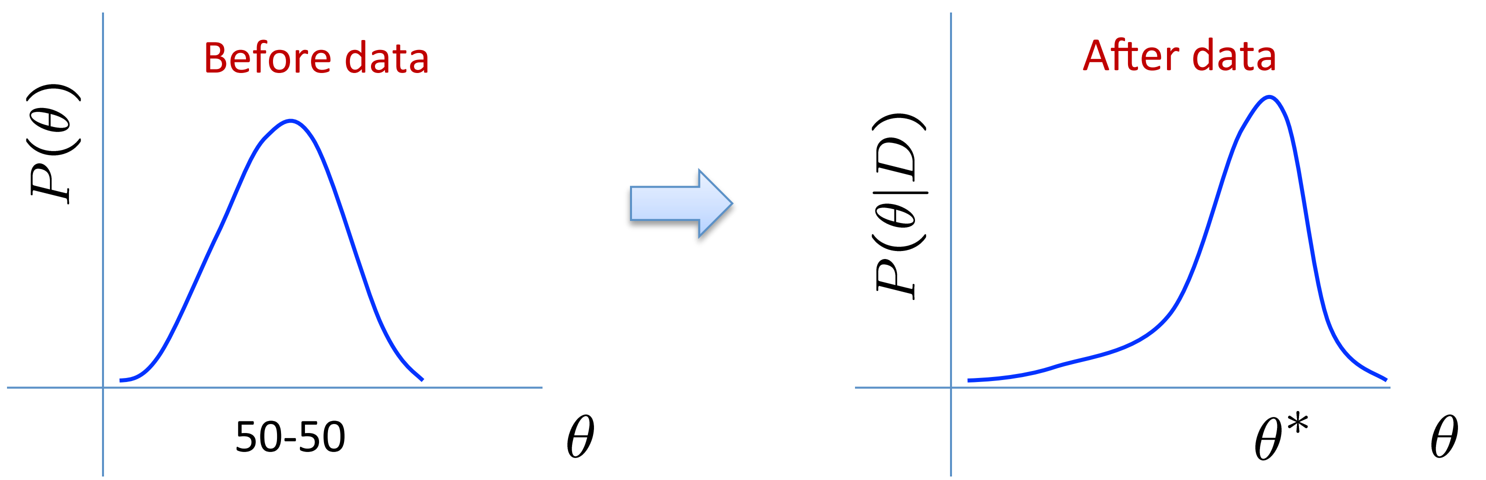
Sample complexity

$$n \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

What about prior knowledge?

- Billionaire says: Wait, I know that the coin is "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:

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

- Or equivalently:

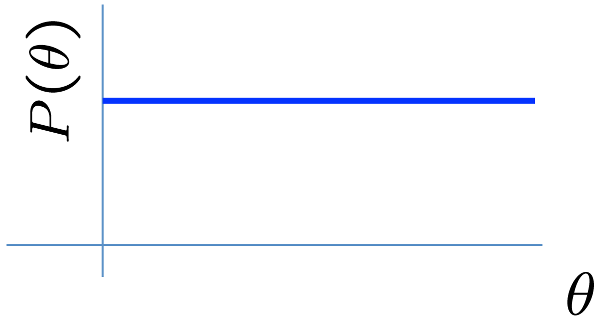
$$\underbrace{P(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{P(\mathcal{D} \mid \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}}$$

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**

Prior distribution

- What about prior?
  - Represents expert knowledge (philosophical approach)
  - Simple posterior form (engineer's approach)

- Uninformative priors:
  - Uniform distribution



- Conjugate priors:
  - Closed-form representation of posterior
  - $P(\theta)$ and $P(\theta|D)$ have the same form

Conjugate Prior

- P(θ) and P(θ|D) have the same form

Eg. 1 Coin flip problem

Likelihood is ~ Binomial

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

If prior is Beta distribution,

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

Then posterior is Beta distribution

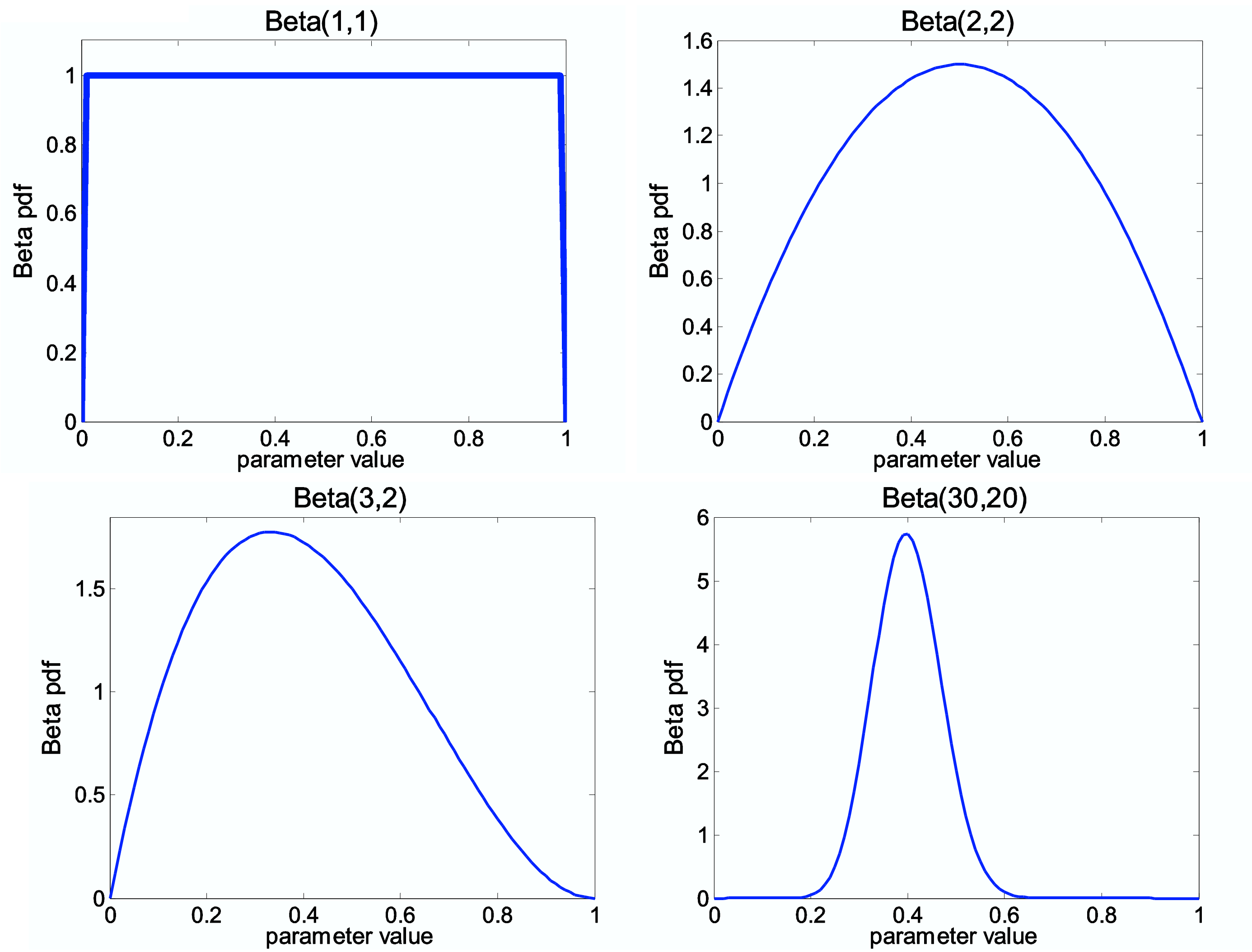
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

For Binomial, conjugate prior is Beta distribution.

Beta distribution

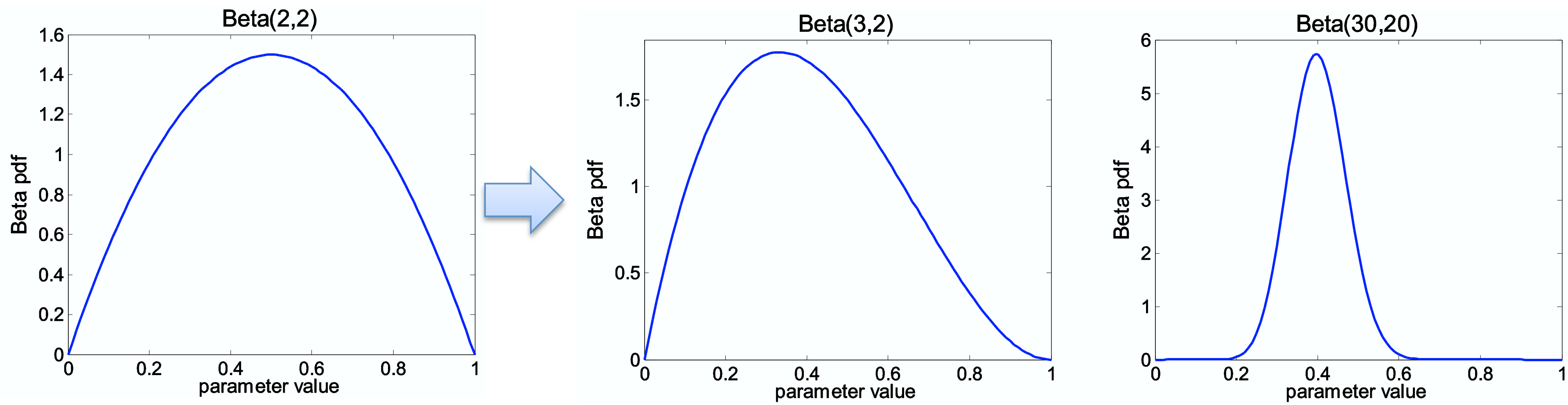
$Beta(\beta_H, \beta_T)$

More concentrated as values of β_H, β_T increase



Beta conjugate prior

$$P(\theta) \sim Beta(\beta_H, \beta_T) \qquad P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



As $n = \alpha_H + \alpha_T$ increases

As we get more samples, effect of prior is "washed out"

Conjugate Prior

- P(θ) and P(θ|D) have the same form

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \frac{n!}{\alpha_1!\alpha_2!\ldots\alpha_k!}\theta_1^{\alpha_1}\theta_2^{\alpha_2}\ldots\theta_k^{\alpha_k} \qquad \sum_{i=1}^{k}\alpha_i = n \qquad \sum_{i=1}^{k}\theta_i = 1$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k}\theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

Maximum A Posteriori Estimation

Choose θ that maximizes a posterior probability

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} \quad P(\theta \mid D) \\ &= \arg\max_{\theta} \quad P(D \mid \theta)P(\theta)\end{aligned}$$

MAP estimate of probability of head:

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$\widehat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

Mode of Beta distribution

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

When is MAP same as MLE?

MLE vs. MAP

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

What if we toss the coin too few times?

- You say: Probability next toss is a head = 0
- Billionaire says: You're fired! ...with prob 1 ☺

$$\widehat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

- Beta prior equivalent to extra coin flips
- As $n \longrightarrow 1$, prior is "forgotten"
- **But, for small sample size, prior is important!**

Bayesians vs. Frequentists

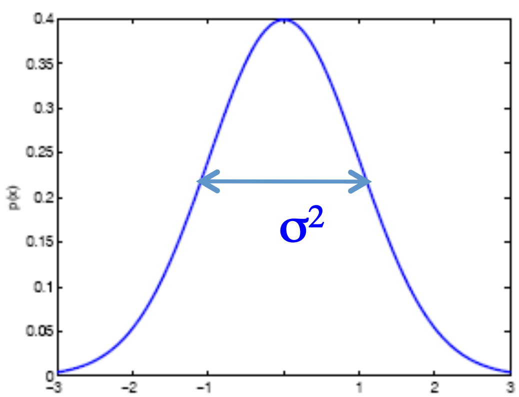
You are no good when sample is small

You give a different answer for different priors

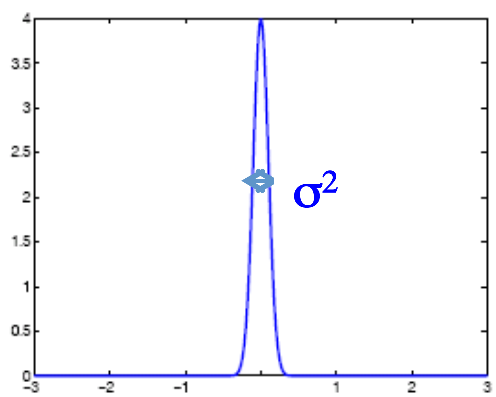
What about continuous variables?

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

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

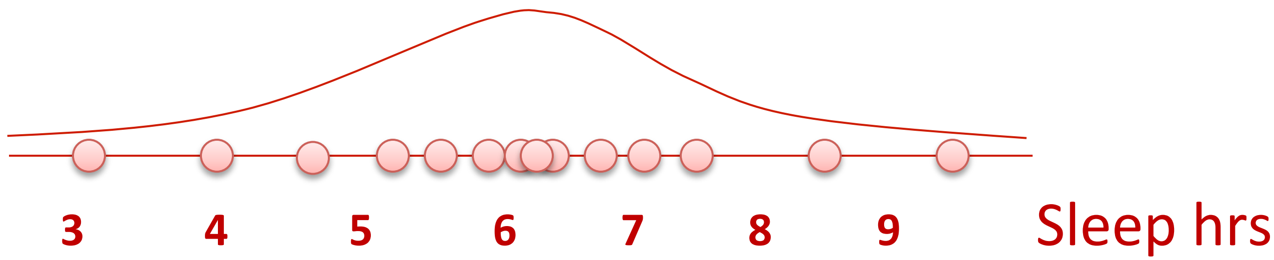


μ=0



μ=0

Gaussian distribution

Data, D = 

- Parameters: μ – mean, σ^2 - variance
- Sleep hrs are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

Properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \; ! \; Y \sim N(a\mu+b,a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2{}_X)$
  - $Y \sim N(\mu_Y,\sigma^2{}_Y)$
  - $Z = X+Y \; ! \; Z \sim N(\mu_X+\mu_Y, \sigma^2{}_X+\sigma^2{}_Y)$

MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-(X_i-\mu)^2/2\sigma^2}$$

Identically distributed

$$= \arg \max_{\theta=(\mu,\sigma^2)} \frac{1}{2\sigma^2} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}$$

MLE for Gaussian mean and variance

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

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

Note: MLE for the variance of a Gaussian is **biased**

– Expected result of estimation is **not** true parameter!

– Unbiased variance estimator:

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

MAP for Gaussian mean and variance

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

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

MAP for Gaussian Mean

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

$$\widehat{\mu}_{MAP} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\eta}{\lambda^2}}{\frac{n}{\sigma^2} + \frac{1}{\lambda^2}}$$

(Assuming known variance σ^2)

Independent of σ^2 if $\lambda^2 = \sigma^2/s$

MAP under Gauss-Wishart prior - Recitation

What you should know…

- Learning parametric distributions: form known, parameters unknown
  - Bernoulli (θ, probability of flip)
  - Gaussian (μ, mean and σ^2, variance)
- MLE
- MAP