Machine Learning

CSE 6363 (Fall 2019)

Lecture 7 MLE MAP

Dajiang Zhu, Ph.D.

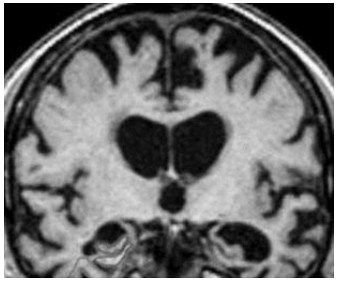
Department of Computer Science and Engineering

Slides of this course (CSE6363) courtesy: Dr. Heng Huang, Dr. Aarti Singh

Performance Measure

Loss and Risk

$\text{loss}(Y, f(X))$ - Measure of closeness between true label Y and prediction $f(X)$

X	Y	$f(X)$	$\text{loss}(Y, f(X))$
	Alzhermer's Disease	Alzhermer's Disease	**0**
		Healthy normal	**1**

$$\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$$ **0 / 1 loss**

Ref: Carlos Guestrin

Performance Measure

Loss and Risk

$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

X	Y	$f(X)$	$\text{loss}(Y, f(X))$
Attribute Information: 1. sepal length in cm 2. sepal width in cm 3. petal length in cm 4. petal width in cm	0.2	0 -- Iris Setosa 1 -- Iris Versicolour 2 -- Iris Virginica	0 ? 1 ? 2 ?

$$\boxed{\text{loss}(Y, f(X)) = (f(X) - Y)^2 \quad \textbf{Square loss}}$$

Ref: Carlos Guestrin

Performance Measure

Loss and Risk

$\text{loss}(Y, f(X))$ - Measure of closeness between true label Y and prediction $f(X)$

Given a brain T1 image drawn randomly from a collection of multiple brain images, how well does the predictor perform on average?

$$\boxed{\text{Risk } R(f) \equiv \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]}$$

Ref: Carlos Guestrin

Performance Measure

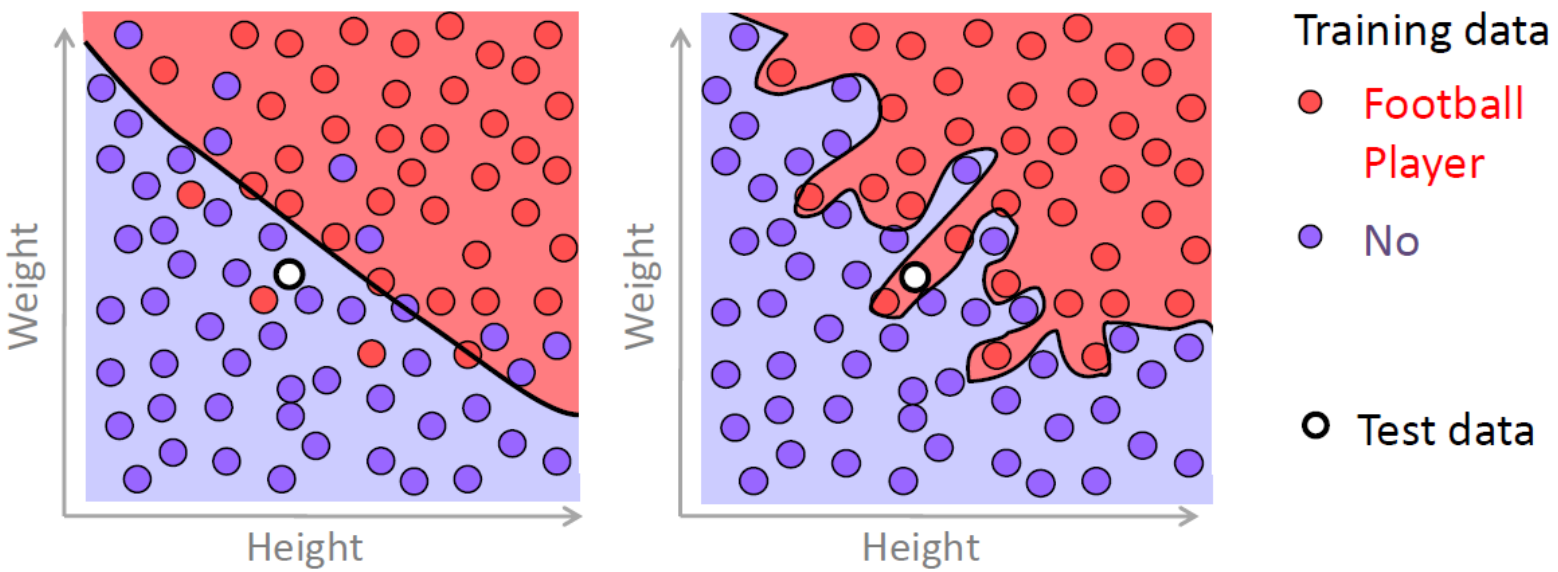
$$\text{Risk } R(f) \equiv \mathbb{E}_{XY}[\text{loss}(Y, f(X))]$$

$\text{loss}(Y, f(X))$	$\text{Risk } R(f)$
$1_{\{f(X)\neq Y\}}$ **0 / 1 loss**	$P(f(X) \neq Y)$ **Probability of Error**
$(f(X) - Y)^2$ **Square loss**	$\mathbb{E}[(f(X) - Y)^2]$ **Mean Square Error**

Ref: Carlos Guestrin

Issues in ML

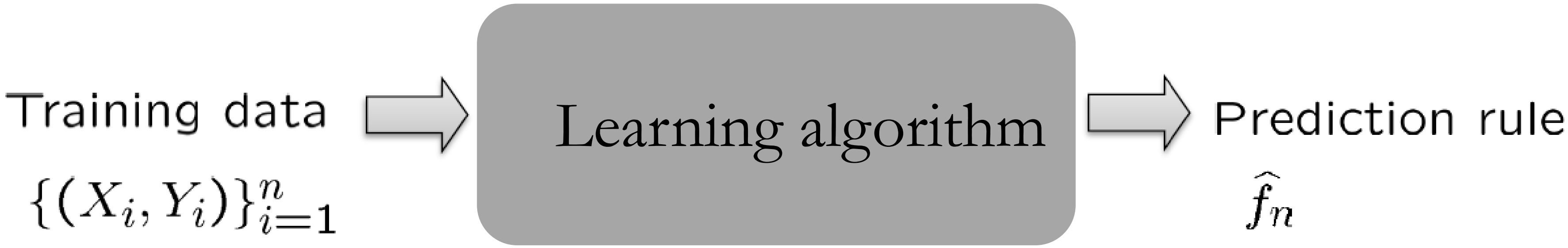
- A good machine learning algorithm
  - Does not **overfit** training data



  - **Generalizes** well to test data

Issues in ML

- One approach -
  - Split available data into two sets $\{(X_i, Y_i)\}_{i=1}^n$ $\{(X_i', Y_i')\}_{i=1}^n$
  - Training Data – used for training the algorithm



  - Test Data (a.k.a. Validation Data, Hold-out Data) – provides estimate of generalization error

$$\textbf{Test Error} = \frac{1}{n}\sum_{i=1}^{n}\left[\text{loss}(Y_i', \widehat{f}_n(X_i'))\right]$$

Probability Distribution

- Let's start from a question
- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - You say: Please flip it a few times

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

Ref: Carlos Guestrin

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1 - θ

- Flips are:
  - Independent events
  - Identically distributed according to Binomial distribution

Ref: Carlos Guestrin

Thumbtack – Binomial Distribution

A binomial experiment is one that possesses the following properties:

- The experiment consists of n repeated trials;
- Each trial results in an outcome that may be classified as a success or a failure (binomial);
- The probability of a success, denoted by p, remains constant from trial to trial and repeated trials are independent.
- The number of successes X in n trials of a binomial experiment is called a binomial random variable.
- The probability distribution of the random variable X is called a binomial distribution, and is given by the formula:

$$P(x) = \binom{n}{k} p^k q^{n-k}$$

n: the number of trials
k:0-n
p: the probability of success in a single trial
q: the probability of failure in a single trial, usually 1-p
$\binom{n}{k}$: combination

If n=1, q=1-p

$$\boxed{P(x) = p^k (1-p)^{1-k}}$$

Bernoulli distribution

Ref: Carlos Guestrin

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1 - θ

- Flips are:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence D of α_H Heads and α_T Tails

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

Ref: Carlos Guestrin

Two Strategies

- Maximum Likelihood Estimation (MLE)
  - Maximizes the probability of observed data

- Maximum A Posteriori Estimation (MAP)
  - Maximizes a posterior probability

Ref: Carlos Guestrin

Two Strategies

- **Maximum Likelihood Estimation (MLE)**
  - **Maximizes the probability of observed data**

- Maximum A Posteriori Estimation (MAP)
  - Maximizes a posterior probability

Ref: Carlos Guestrin

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function?
- MLE: Choose θ that maximizes the probability of observed data:

$$
\begin{aligned}
\widehat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\
&= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta)
\end{aligned}
$$

Ref: Carlos Guestrin

Your Second Learning Algorithm

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

How?

- Set derivative to zero:

$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0$$

Ref: Carlos Guestrin

Simple bound

Based on 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 $\varepsilon > 0$:

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

Ref: Carlos Guestrin

Simple bound

Imaging we are facing the following problem:

- Billionaire says: I want to know the coin parameter Θ, within ε= 0.1, with probability at least 1-δ= 0.95.
  How many flips?

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

Ref: Carlos Guestrin

PAC Learning

- PAC: Probably Approximate Correct

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

Ref: Carlos Guestrin

What about Prior

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

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: $\theta = 3/5$, I can prove it!
- Billionaire says: Wait, I know that the thumbtack is "close" to 50-50. What can you do?
- **You say: I can learn it the Bayesian way…**
- Rather than estimating a single θ, we obtain a distribution over possible values of θ

Ref: Carlos Guestrin

Bayesian Learning

- Use Bayes rule:

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

- Or equivalently:

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

Ref: Carlos Guestrin

Bayesian Learning for Thumbtack

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

- Likelihood function is simply Binomial:

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

- What about prior?
  - Represent expert knowledge
  - Simple posterior form

Prior -> observe the data->Posterior

Ref: Carlos Guestrin

Bayesian Learning for Thumbtack

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

- Conjugate priors:
  - Closed-form representation of posterior
  - **P(θ) and P(θ|D) have the same form**

Ref: Carlos Guestrin

Bayesian Learning for Thumbtack

P(θ) and P(θ|D) have the same form?

- Back to coin flip problem

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

✓ Likelihood is ~ Binomial

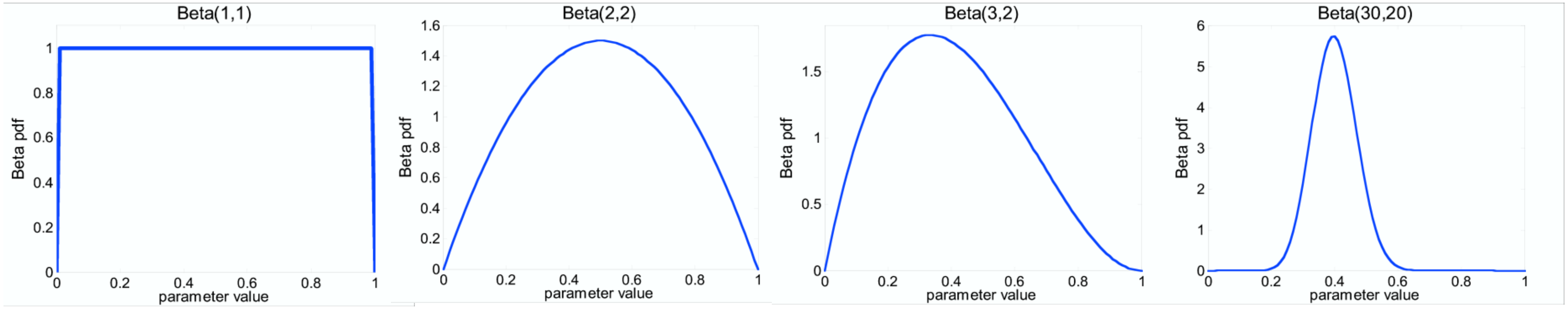
✓ If prior is Beta distribution:

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

Ref: Carlos Guestrin

Beta Prior Distribution – P(θ)

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



More concentrated as values of β_H, β_T increase

Ref: Carlos Guestrin

Posterior Distribution

- Prior: $Beta(\beta_H, \beta_T)$

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

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

- Data: α_H Heads and α_T Tails
- Posterior distribution:

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

Ref: Carlos Guestrin

Bayesian Learning for Thumbtack

P(θ) and P(θ|D) have the same form?

- Back to coin flip problem

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

✓ Likelihood is ~ Binomial

✓ If prior is Beta distribution:

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

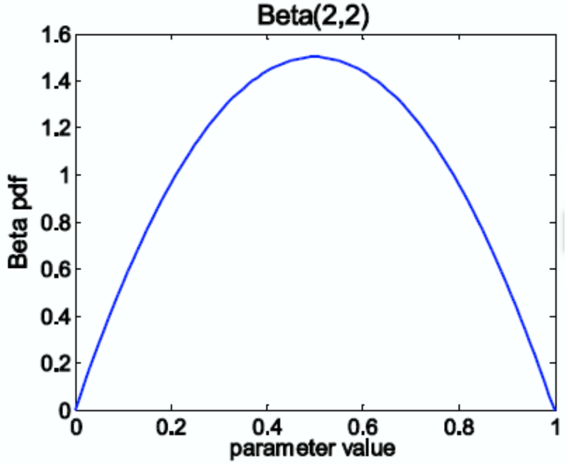
✓ Then posterior is Beta distribution

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

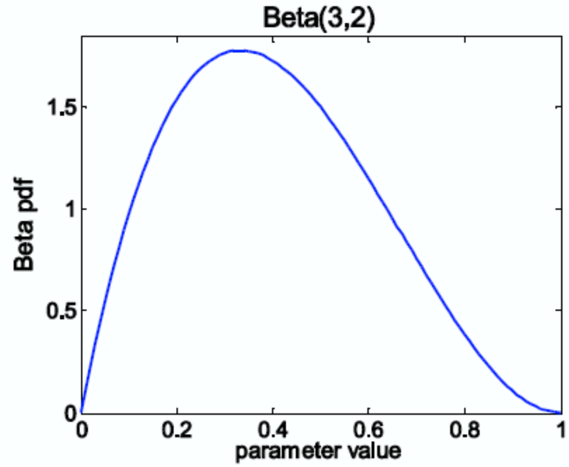
Ref: Carlos Guestrin

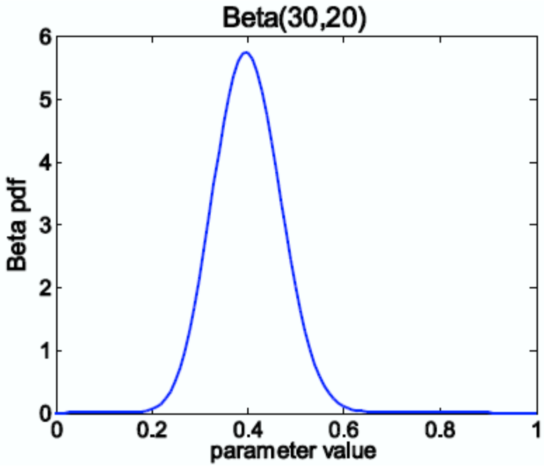
Beta Prior Distribution – P(θ)

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



Prior





…

As we get more samples (↑n=$\alpha_H + \alpha_T$), effect of prior is "washed out"

Ref: Carlos Guestrin

The Beta Distribution

- To ensure the prior is normalized, we define

$$P(\mu|a,b) = \text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

  where the gamma function is defined as

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$$

  Note that $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1) = 1$. Also, for integers, $\Gamma(x+1) = x!$.

- The normalization constant $1/Z(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ ensures

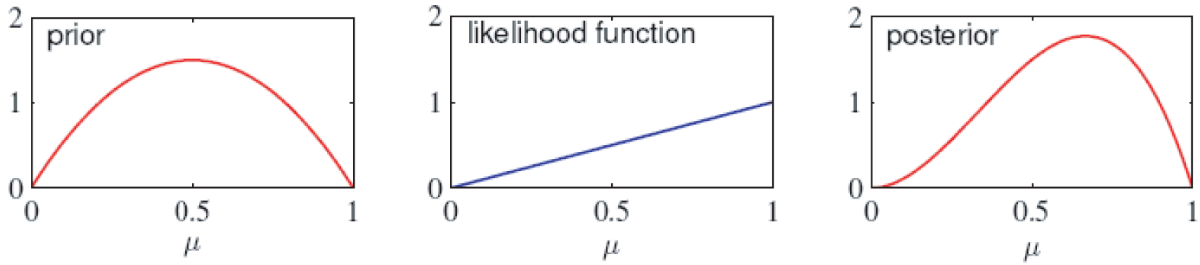
$$\int_0^1 \text{Beta}(\mu|a,b)d\mu = 1$$

Ref: Kevin Murphy

Bayesian Updating in Pictures

- Start with $Be(\mu|a = 2, b = 2)$ and observe $x = 1$, so the posterior is $Be(\mu|a = 3, b = 2)$.

```
thetas = 0:0.01:1;
alphaH = 2; alphaT = 2; Nh=1; Nt=0; N = Nh+Nt;
prior = betapdf(thetas, alphaH, alphaT);
lik = choose(N,Nh) * thetas.^Nh .* (1-thetas).^Nt;
post = betapdf(thetas, alphaH+Nh, alphaT+Nt);
```



Ref: Kevin Murphy

Effect of Prior Strength

- Let $N = N_h + N_t$ be number of samples (observations).
- Let N' be the number of pseudo observations (strength of prior) and define the prior means

$$\alpha_h = N'\alpha'_h, \quad \alpha_t = N'\alpha'_t, \quad \alpha'_h + \alpha'_t = 1$$

- Then posterior mean is a convex combination of the prior mean and the MLE (where $\lambda = N'/(N+N')$):

$$\begin{aligned}
P(X = h|\alpha_h, \alpha_t, N_h, N_t) &= \frac{\alpha_h + N_h}{\alpha_h + N_h + \alpha_t + N_t} \\
&= \frac{N'\alpha'_h + N_h}{N + N'} \\
&= \frac{N'}{N + N'}\alpha'_h + \frac{N}{N + N'}\frac{N_h}{N} \\
&= \lambda\alpha'_h + (1 - \lambda)\frac{N_h}{N}
\end{aligned}$$

Ref: Kevin Murphy

Effect of Prior Strength

- Suppose we have a uniform prior $\alpha'_h = \alpha'_t = 0.5$, and we observe $N_h = 3$, $N_t = 7$.
- Weak prior $N' = 2$. Posterior prediction:

$$P(X = h|\alpha_h = 1, \alpha_t = 1, N_h = 3, N_t = 7) = \frac{3+1}{3+1+7+1} = \frac{1}{3} \approx 0.33$$

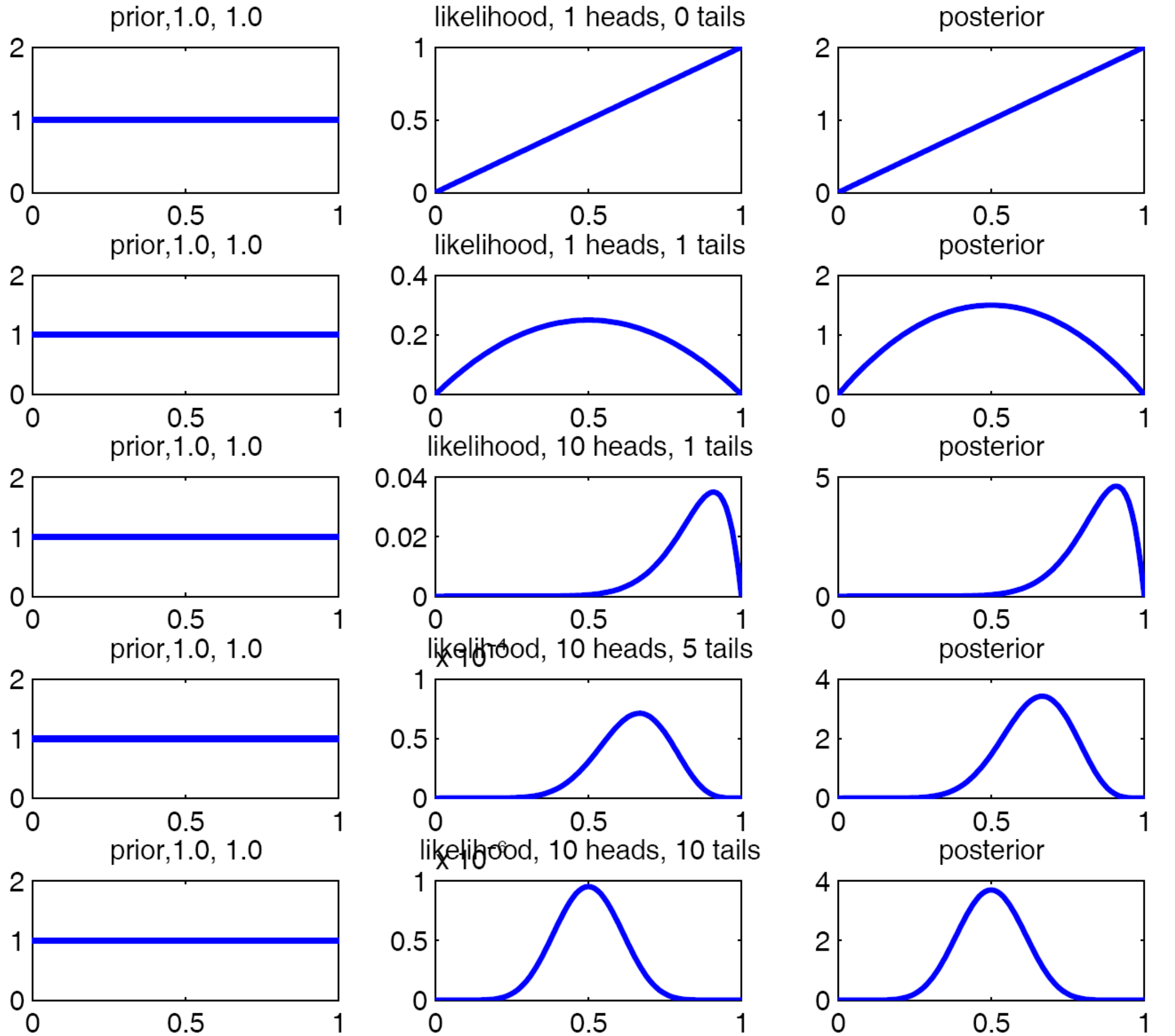
- Strong prior $N' = 20$. Posterior prediction:

$$\frac{3+10}{3+10+7+10} = \frac{13}{30} \approx 0.43$$

- However, if we have enough data, it washes away the prior. e.g., $N_h = 300$, $N_t = 700$. Estimates are $\frac{300+1}{1000+2}$ and $\frac{300+10}{1000+20}$, both of which are close to 0.3
- As $N \to \infty$, $P(\theta|D) \to \delta(\theta, \hat{\theta}_{ML})$, so $E[\theta|D] \to \hat{\theta}_{ML}$.

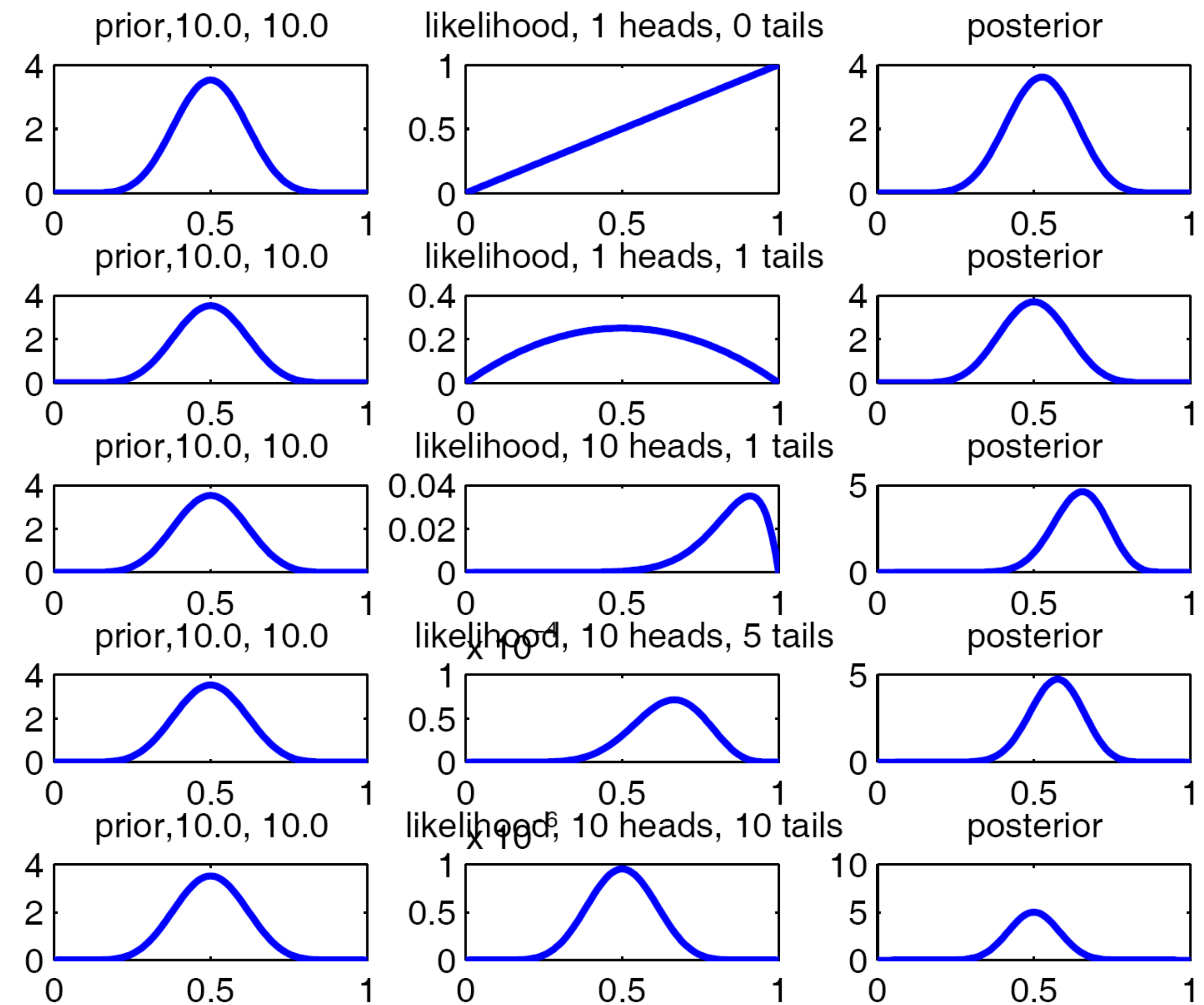
Ref: Kevin Murphy

Parameter Posterior – Small Sample, Uniform Prior



Ref: Kevin Murphy

Parameter Posterior – Small Sample, Strong Prior



Ref: Kevin Murphy

From Coin to Dice

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

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

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

Ref: Kevin Murphy

Two Strategies

- Maximum Likelihood Estimation (MLE)
  - Maximizes the probability of observed data

- **Maximum A Posteriori Estimation (MAP)**
  - **Maximizes a posterior probability**

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

Ref: Carlos Guestrin

Maximum A Posteriori (MAP) Estimation

- MAP estimation picks the mode of the posterior

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(D|\theta)p(\theta)$$

- If $\theta \sim Be(a, b)$, this is just

$$\hat{\theta}_{MAP} = (a-1)/(a+b-2)$$

- MAP is equivalent to maximizing the penalized maximum log-likelihood

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(D|\theta) - \lambda c(\theta)$$

where $c(\theta) = -\log p(\theta)$ is called a *regularization term*. λ is related to the strength of the prior.

Ref: Kevin Murphy

Summarize MLE and 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?

Ref: Kevin Murphy

Summarize MLE and MAP

- What if we only toss the coin 3 times:

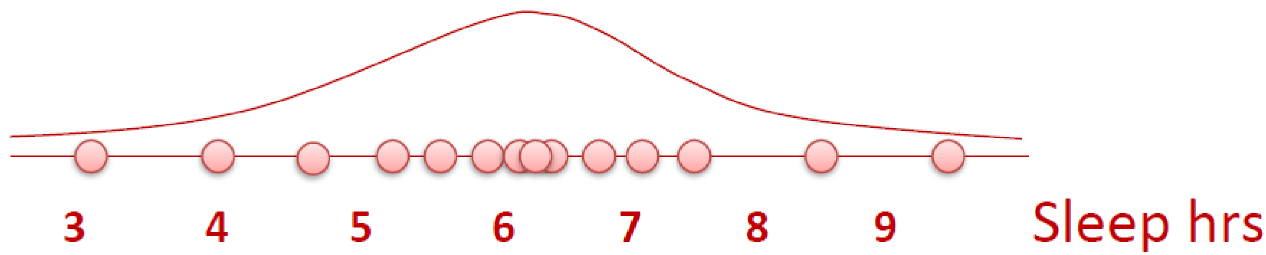
$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

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

- ❖ Beta prior equivalent to extra coin flips **(regularization term)**
- ❖ As n→∞, prior is "forgotten"
- ❖ For small sample size, prior is important

Ref: Kevin Murphy

About Gaussian

Data, D = 

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

Gaussian Density in 1-D

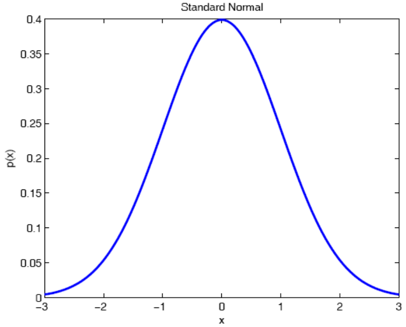
- If $X \sim N(\mu, \sigma^2)$, the probability density function (pdf) of X is defined as

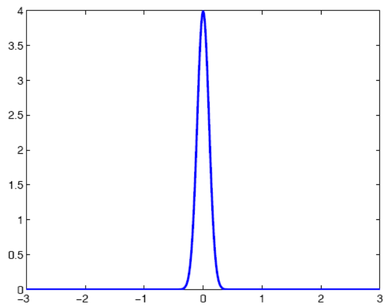
$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

We will often use the precision $\lambda = 1/\sigma^2$ instead of the variance σ^2.

- Note that a density evaluated at a point can be bigger than 1!
- Here is how we plot the pdf in matlab

```
xs=-3:0.01:3;  plot(xs,normpdf(xs,mu,sigma))
```





Ref: Kevin Murphy

Properties of Gaussian

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

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

Properties of Gaussian

- All Gaussians are similar in shape and symmetric
- Within 1 standard deviation of the mean - 68.3%
- Within 2 standard deviation of the mean - 95.45%
- Within 3 standard deviation of the mean - 99.7%
- Full width at half maximum (FWHM) – 2.35 standard deviation

Multivariate Gaussian

1-dimensional Gaussian

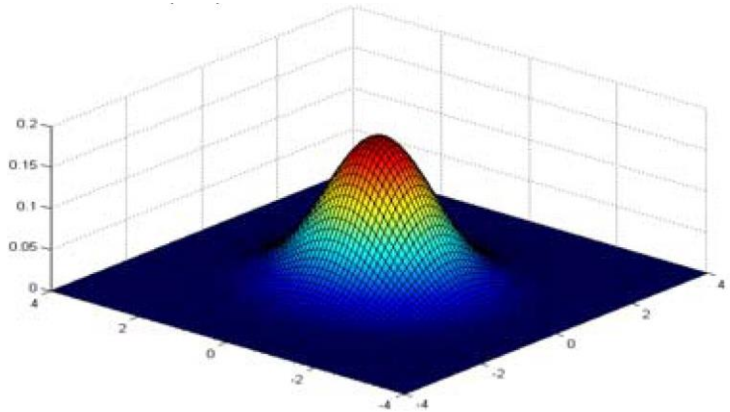
$$p(x|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

2-dimensional Gaussian

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

d-dimensional Gaussian

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



Multivariate Gaussian

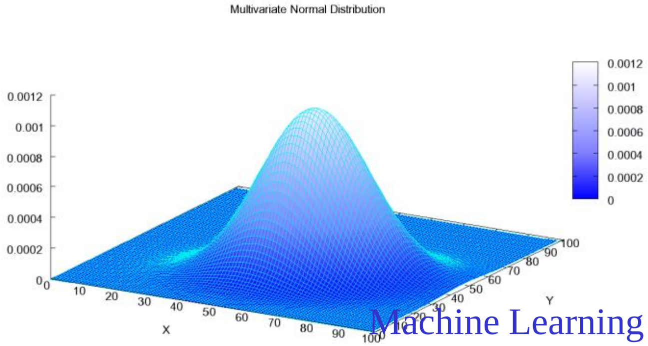
- If $X \in \mathbb{R}^d$ is a jointly gaussian random vector, then its pdf is

$$p(x) = N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

- The quantity $\Delta^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$ is called the Mahalanobis distance between x and μ.
- The first and second moments are

$$E[X] = \mu, \quad \mathrm{Cov}[X] = \Sigma$$

- Sometimes we will use the precision matrix Σ^{-1} instead of the covariance matrix Σ.



Ref: Kevin Murphy