Slides originally taken from
http://research.microsoft.com/en-us/um/people/cmbishop/prml/
and modified by Pushpendre Rastogi for cs475-2017

# PATTERN RECOGNITION AND MACHINE LEARNING

## CHAPTER 2: PROBABILITY DISTRIBUTIONS

## Parametric Distributions

Basic building blocks: $p(\mathbf{x}|\boldsymbol{\theta})$

Need to determine $\boldsymbol{\theta}$ given $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Representation: $\boldsymbol{\theta}^{\star}$ or $p(\boldsymbol{\theta})$ ?

## Binary Variables (1)

Coin flipping: heads=1, tails=0

$$p(x = 1|\mu) = \mu$$

Bernoulli Distribution

$$\begin{aligned} \mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu) \end{aligned}$$

## Binary Variables (1)

Coin flipping: heads=1, tails=0

$$p(x = 1|\mu) = \mu$$

Note the mathematical language

Bernoulli Distribution

$$\begin{aligned} \mathrm{Bern}(x|\mu) &= \boxed{\mu^x(1-\mu)^{1-x}} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu) \end{aligned}$$

## Parameter Estimation (MLE)

- Given $\mathcal{D} = \{x_1, x_2, \dots, x_N\}$ and a set of probability distributions $\{p_\theta | \ \theta \in \Theta\}$ choose the parameter that maximizes $p_\theta(\mathcal{D})$
- Usually we assume $\mathcal{D}$ is drawn *iid* from $p_\theta$ in which case $p_\theta(\mathcal{D}) = \prod_{\{i=1 \ to \ N\}} p_\theta(x_i)$
- **Note** $\text{argmax}_x f(x) = \text{argmax}_x \log(f(x))$

## Parameter Estimation (1)

ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \dots, x_N\}$, $m$ heads $(1)$, $N - m$ tails $(0)$

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

$$\mu_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$$

## Parameter Estimation (2)

Example: $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$

Prediction: *all* future tosses will land heads up

# Overfitting to $\mathrm{D}$

## Parameter Estimation (2)

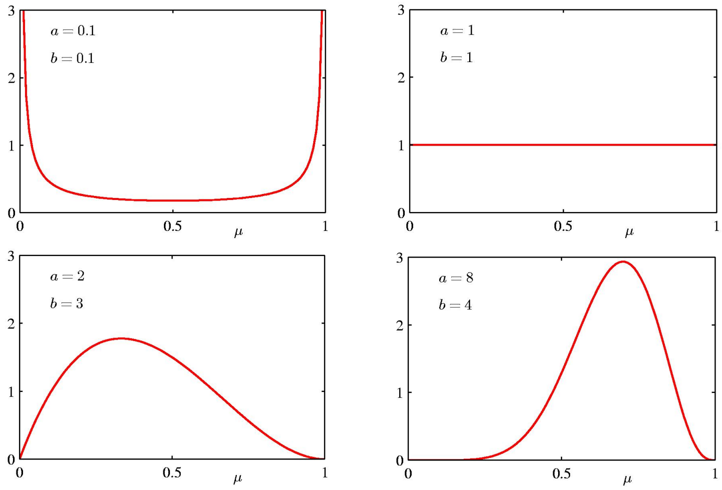
Example: $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$

Prediction: *all* future tosses will land heads up

# Overfitting to $\mathrm{D}$


Let's try a different approach.

# Beta Distribution




# Beta Distribution

Distribution over $\mu \in [0, 1]$.

$$
\begin{aligned}
\mathrm{Beta}(\mu|a, b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \\
\mathbb{E}[\mu] &= \frac{a}{a+b} \\
\mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}
$$

## Beta Distribution

Distribution over $\mu \in [0, 1]$.

$$\begin{aligned}
\mathrm{Beta}(\mu|a,b) &= \boxed{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}\mu^{a-1}(1 \\
\mathbb{E}[\mu] &= \frac{a}{a+b} \\
\mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}$$

For simplicity
Γ(a) = factorial(a-1)

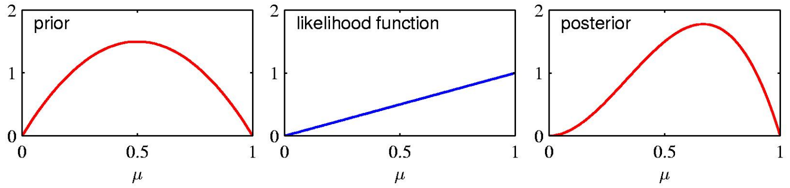
## Bayesian Bernoulli

$$\begin{aligned}
p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\
&= \left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}\right) \mathrm{Beta}(\mu|a_0, b_0) \\
&\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \\
&\propto \mathrm{Beta}(\mu|a_N, b_N)
\end{aligned}$$

$$a_N = a_0 + m \qquad b_N = b_0 + (N - m)$$

1. The Beta distribution provides the *conjugate* prior for the Bernoulli distribution.

# Prior · Likelihood = Posterior




# Bayesian Bernoulli

$$\begin{aligned} p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\ &= \left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}\right) \mathrm{Beta}(\mu|a_0, b_0) \\ &\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \\ &\propto \mathrm{Beta}(\mu|a_N, b_N) \end{aligned}$$

$$a_N = a_0 + m \qquad b_N = b_0 + (N-m)$$

1. The Beta distribution provides the *conjugate* prior for the Bernoulli distribution.
2. If we chose μ that maximizes Beta($\mu \mid a_N, b_N$) and use that for all subsequent predictions then we are performing MAP Estimation.