Machine Learning

4771

Instructors:
Adrian Weller and Ilia Vovsha

Topic 2: Basic concepts of Bayesians and Frequentists

- Properties of PDFs
- Bayesians & Frequentists
- ML, MAP and Full Bayes
- Example: Coin Toss
- Bernoulli Priors
- Conjugate Priors
- Bayesian decision theory

Properties of PDFs

Probability
Distribution
Function

- Review some basics of probability theory
- First, pdf is a function, multiple inputs, one output:

$p\left(x_1,\ldots,x_n\right)$ $\qquad p\left(X_1 = 0.3,\ldots,X_n = 1\right) = 0.2$

- Function's output is always non-negative:

$p\left(x_1,\ldots,x_n\right) \geq 0$

- Can have discrete or continuous inputs or both:

$p\left(X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 3.1415\right)$

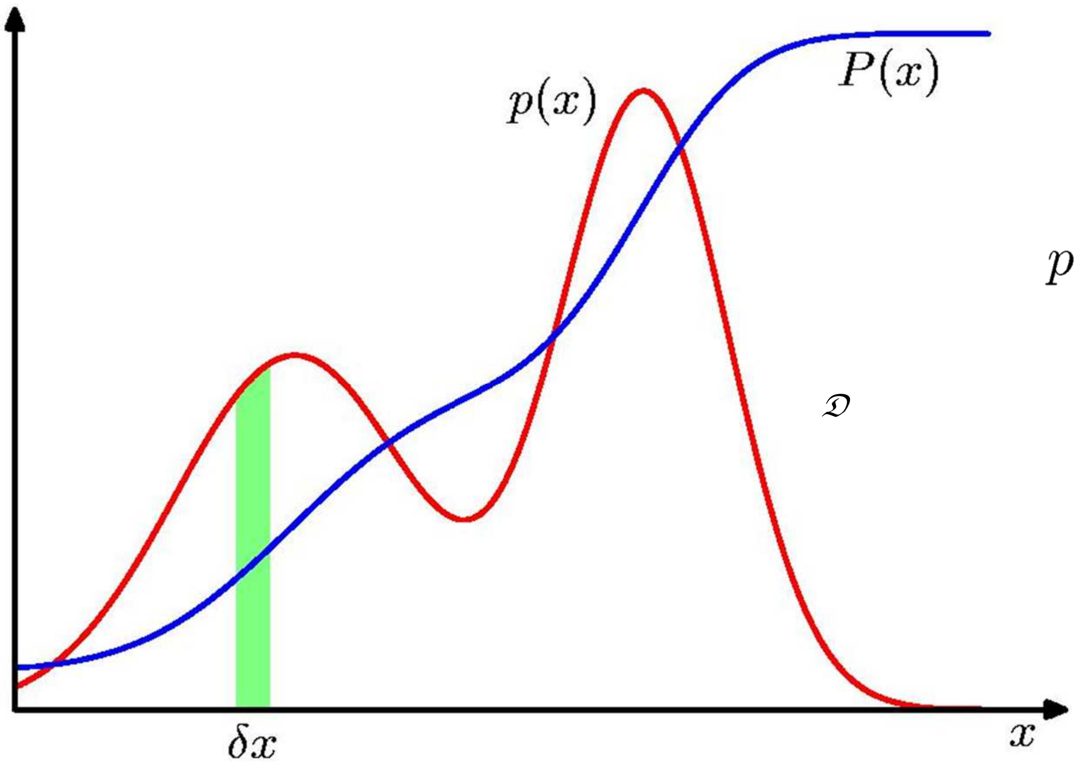
- Summing over the domain of all inputs gives unity:

$\int_{y=-\infty}^{\infty}\int_{x=-\infty}^{\infty} p\left(x,y\right)dxdy = 1$ $\qquad \sum_y \sum_x p\left(x,y\right) = 1$

Y \ X	0	1
0	0.4	0.1
1	0.3	0.2

Continuous→integral, Discrete→sum

Properties of PDFs



PDF

$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

CDF

Cumulative Distribution Function

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Y \ X	0	1
0	0.4	0.1
1	0.3	0.2

Properties of PDFs

- Marginalizing: integrate/sum out a variable leaves a marginal distribution over the remaining ones…

$$\sum_y p(x,y) = p(x)$$

- Conditioning: if a variable 'y' is 'given' we get a conditional distribution over the remaining ones…

$$p(x \mid y) = \frac{p(x,y)}{p(y)}$$

- Bayes Rule: mathematically just redo conditioning but has a deeper meaning (1764)… if we have x being data and θ being a model

likelihood

posterior

$$p(\theta \mid x) = \frac{p(x \mid \theta)\,p(\theta)}{p(x)}$$

prior

evidence

Properties of PDFs

- Expectation: can use pdf p(x) to compute averages and expected values for quantities, denoted by:

$$E_{p(x)}\left\{f(x)\right\} = \int_x p(x) f(x) dx \quad or \quad = \sum_x p(x) f(x)$$

- Properties: $E\left\{cf(x)\right\} = cE\left\{f(x)\right\}$

$$E\left\{f(x) + c\right\} = E\left\{f(x)\right\} + c$$

$$E\left\{E\left\{f(x)\right\}\right\} = E\left\{f(x)\right\}$$

- Mean: expected value for x

$$E_{p(x)}\left\{x\right\} = \int_{-\infty}^{\infty} p(x) x \, dx$$

example: speeding ticket

Fine=0$	Fine=20$
0.8	0.2

expected cost of speeding?

- Variance: expected value of (x-mean)^2, how much x varies

$$Var\left\{x\right\} = E\left[\left(x - E\left\{x\right\}\right)^2\right] = E\left\{x^2 - 2xE\left\{x\right\} + E\left\{x\right\}^2\right\}$$

$$= E\left\{x^2\right\} - 2E\left\{x\right\}E\left\{x\right\} + E\left\{x\right\}^2 = E\left\{x^2\right\} - E\left\{x\right\}^2$$

The IID Assumption

- Most of the time, we will assume that a dataset is independent and identically distributed (IID)

- In many real situations, data is generated by some black box phenomenon in an arbitrary order.
- Assume we are given a dataset:

$$\mathscr{D} = \left\{ x_1, \ldots, x_N \right\}$$

"Independent" means that (given the model θ) the probability of our data multiplies:

$$p\left(x_1, \ldots, x_N \mid \Theta\right) = \prod_{i=1}^{N} p_i\left(x_i \mid \Theta\right)$$

"Identically distributed" means that each marginal probability is the same for each data point

$$p\left(x_1, \ldots, x_N \mid \Theta\right) = \prod_{i=1}^{N} p_i\left(x_i \mid \Theta\right) = \prod_{i=1}^{N} p\left(x_i \mid \Theta\right)$$

Ex: Is a coin fair?

A stranger tells you his coin is fair.

Let's assume tosses are iid with $P(H)=\mu$.

He tosses it 4 times, gets H H T H.

What can you say about μ?

Bayesians & Frequentists

•Frequentists (Neymann/Pearson/Wald). An orthodox view that sampling is infinite and decision rules can be sharp.

•Bayesians (Bayes/Laplace/de Finetti). Unknown quantities are treated probabilistically and the state of the world can always be updated.

actuarial fair

de Finetti: p(event) = price I would pay for a contract that pays $1 when event happens

•Likelihoodists (Fisher). Single sample inference based on maximizing the likelihood function.

Bayesians & Frequentists

- Frequentists:
  - Data are a repeatable random sample - there is a frequency
  - Underlying parameters remain constant during this repeatable process
  - Parameters are fixed

- Bayesians:
  - Data are observed from the realized sample.
  - Parameters are unknown and described probabilistically
  - Data are fixed

Frequentists

- Frequentists: classical / objective view / no priors
  every statistician should compute same p(x) so no priors
  can't have a p(event) if it never happened
  avoid p(θ), there is 1 true model, not distribution of them
  permitted: $p_\theta(x,y)$ forbidden: $p(x,y|\theta)$
  Frequentist inference: estimate one best model θ
    use the Maximum Likelihood Estimator (ML)
    (unbiased & minimum variance)
    do not depend on Bayes rule for learning

$$\theta_{ML} = \arg\max_\theta p\left(\mathcal{D} \mid \theta\right)$$

$$\text{Data } \mathcal{D} = \left(x_1, x_2, \ldots, x_n\right)$$

Bayesians

- Bayesians: subjective view / priors are ok
  put a distribution or pdf on all variables in the problem
  even models & deterministic quantities (speed of light)
  use a prior $p(\theta)$ on the model θ before seeing any data

  Bayesian inference: use Bayes rule for learning, integrate over all model (θ) unknown variables

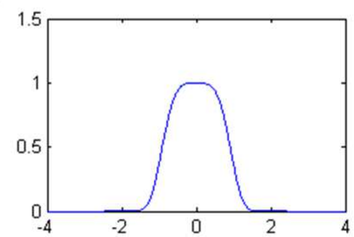
Bayesian Inference

- Bayes rule can lead us to maximum likelihood
- Assume we have a prior over models p(θ)

$$\underbrace{p(\theta \mid x)}_{\text{posterior}} = \frac{\overbrace{p(x \mid \theta)}^{\text{likelihood}}\,\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(x)}_{\text{evidence}}}$$

- How to pick p(θ)?
  Pick simpler θ is better
  Pick form for mathematical convenience

$$p(\theta)$$



- We have data (can assume IID): $\mathcal{D} = \left\{x_1, x_2, \ldots, x_N\right\}$
- Want to get a model to compute: $p(x)$
- Want p(x) given our data… How to proceed?

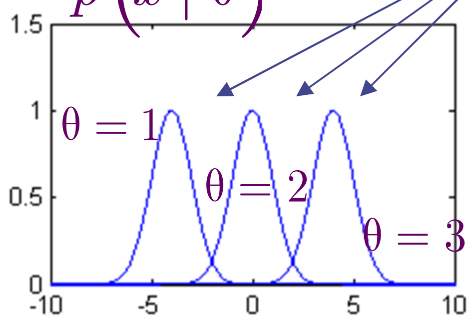
Bayesian Inference

- Want p(x) given our data… $p\left(x \mid \mathcal{D}\right) = p\left(x \mid x_1, x_2, \ldots, x_n\right)$

$$
\begin{aligned}
p\left(x \mid \mathcal{D}\right) &= \int_\theta p\left(x, \theta \mid \mathcal{D}\right) d\theta \\
&= \int_\theta p\left(x \mid \theta, \mathcal{D}\right) p\left(\theta \mid \mathcal{D}\right) d\theta \\
&= \int_\theta p\left(x \mid \theta, \mathcal{D}\right) \frac{p\left(\mathcal{D} \mid \theta\right) p\left(\theta\right)}{p\left(\mathcal{D}\right)} d\theta \\
&= \int_\theta p\left(x \mid \theta\right) \frac{\prod_{i=1}^{N} p\left(x_i \mid \theta\right) p\left(\theta\right)}{p\left(\mathcal{D}\right)} d\theta
\end{aligned}
$$

Prior

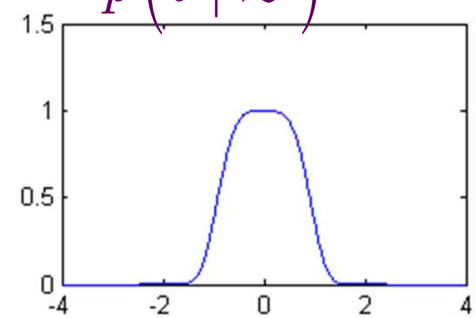
$p\left(x \mid \theta\right)$



Many models

Weight on each model

$p\left(\theta \mid \mathcal{D}\right)$



Bayesian Inference to MAP & ML

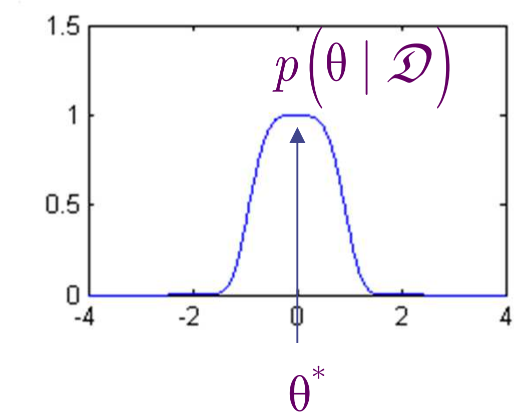
- The full Bayesian Inference integral can be mathematically tricky. MAP and ML are approximations of it…

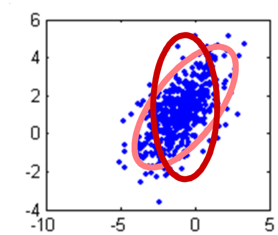
$$p(x \mid \mathcal{D}) = \int_\theta p(x \mid \theta) \frac{\prod_{i=1}^N p(x_i \mid \theta) p(\theta)}{p(\mathcal{D})} d\theta$$

$$\approx \int_\theta p(x \mid \theta) \delta(\theta - \theta^*) d\theta$$

$$where\ \theta^* = \begin{cases} \arg\max_\theta \dfrac{\prod_{i=1}^N p(x_i \mid \theta) p(\theta)}{p(\mathcal{D})} & MAP \\ \arg\max_\theta \dfrac{\prod_{i=1}^N p(x_i \mid \theta) uniform(\theta)}{p(\mathcal{D})} & ML \end{cases}$$

$p(\theta \mid \mathcal{D})$





- Maximum A Posteriori (MAP) is like Maximum Likelihood (ML) with a prior p(θ) which lets us prefer some models over others

$$l_{MAP}(\theta) = l_{ML}(\theta) + \log p(\theta) = \sum_{i=1}^N \log p(x_i \mid \theta) + \log p(\theta)$$

Ex: Is a coin fair?

A stranger tells you his coin is fair.

Let's assume tosses are iid with P(H)=μ.

He tosses it 4 times, gets H H T H. $\mathcal{D} = (H, H, T, H)$

What can you say about μ?

Bernoulli Probability ML

0 ~ Tail
1 ~ Head
μ=P(H)

- Bernoulli: $p(x) = \mu^{x}(1-\mu)^{1-x} \quad \mu \in [0,1] \quad x \in \{0,1\}$
- Log-Likelihood (IID): $\sum_{i=1}^{N} \log p(x_i \mid \mu) = \sum_{i=1}^{N} \log \mu^{x_i}(1-\mu)^{1-x_i}$
- Gradient=0:

$$\frac{\partial}{\partial\mu}\sum_{i=1}^{N} \log \mu^{x_i}(1-\mu)^{1-x_i} = 0$$

$$\frac{\partial}{\partial\mu}\sum_{i=1}^{N} x_i \log\mu + (1-x_i)\log(1-\mu) = 0$$

$$\frac{\partial}{\partial\mu}\sum_{i\in class1} \log\mu + \sum_{i\in class0}\log(1-\mu) = 0$$

$$\sum_{i\in class1}\tfrac{1}{\mu} - \sum_{i\in class0}\tfrac{1}{1-\mu} = 0$$

$$m\tfrac{1}{\mu} - (N-m)\tfrac{1}{1-\mu} = 0$$

$$m(1-\mu) - (N-m)\mu = 0$$

$$m - N\mu = 0$$

$$\mu = \tfrac{m}{N}$$

N *trials/tosses*
m *heads/1s*
N-m *tails/0s*

x=0	x=1
$\frac{N-m}{N}$	$\frac{m}{N}$

Bernoulli Bayes, Prior 1

$$\mathcal{D} = (H, H, T, H)$$

- Assume prior μ=1/2, point mass distribution
- Posterior

$$P(\mu = r \mid \mathcal{D}) \propto P(\mathcal{D} \mid \mu = r) \times P(\mu = r)$$

- If the prior is 0 for some value, the posterior will also be 0 at that value no matter what data we see

Bernoulli Bayes, Prior 2

$$\mathcal{D} = (H, H, T, H)$$

- Allow some chance of bias
- Prior P(μ=1/2) = 1-b
  P(μ=3/4) = b
- Posterior

$$P(\mu = \frac{1}{2} \mid \mathcal{D}) \propto P(\mathcal{D} \mid \mu = \frac{1}{2}) \times P(\mu = \frac{1}{2}) = \frac{1}{2^4} \times (1-b)$$

$$P(\mu = \frac{3}{4} \mid \mathcal{D}) \propto P(\mathcal{D} \mid \mu = \frac{3}{4}) \times P(\mu = \frac{3}{4}) = \left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right) \times b$$

Bernoulli Bayes, Prior 2

$$\mathcal{D} = (H, H, T, H)$$

- Prior $\quad P(\mu=1/2) = 1\text{-}b$

  $\quad\quad\quad P(\mu=3/4) = b$

- Posterior

$$P(\mu = \tfrac{1}{2} \mid \mathcal{D}) \propto P(\mathcal{D} \mid \mu = \tfrac{1}{2}) \times P(\mu = \tfrac{1}{2}) = \frac{1}{2^4} \times (1-b)$$

$$P(\mu = \tfrac{3}{4} \mid \mathcal{D}) \propto P(\mathcal{D} \mid \mu = \tfrac{3}{4}) \times P(\mu = \tfrac{3}{4}) = \left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right) \times b$$

The two are equal when $b = \dfrac{16}{43}$

Bernoulli Bayes, Prior 3 $\mathcal{D} = (H, H, T, H)$

- Uniform prior $\mu \sim U[0,1]$

$$P(\mu = r \mid \mathcal{D}) \propto P(\mathcal{D} \mid \mu = r) \times P(\mu = r)$$

$$P(\mathcal{D} \mid \mu = r) = r^3(1-r)$$

$$\int_0^1 r^3(1-r)dr = \frac{1}{4} - \frac{1}{5} = \frac{1}{20}$$

Where's the mode?

Hence posterior $P(\mu = r \mid \mathcal{D}) = 20r^3(1-r)$

Notice for N tosses with m heads, $N-m$ tails,

$$P(\mathcal{D} \mid \mu = r) = r^m(1-r)^{N-m}$$

Does this suggest a convenient prior?

Beta Distribution

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

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

Here switch to typical ‘sloppy’ notation, using µ for variable name and its value.

Bernoulli Bayes, Beta Prior

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

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

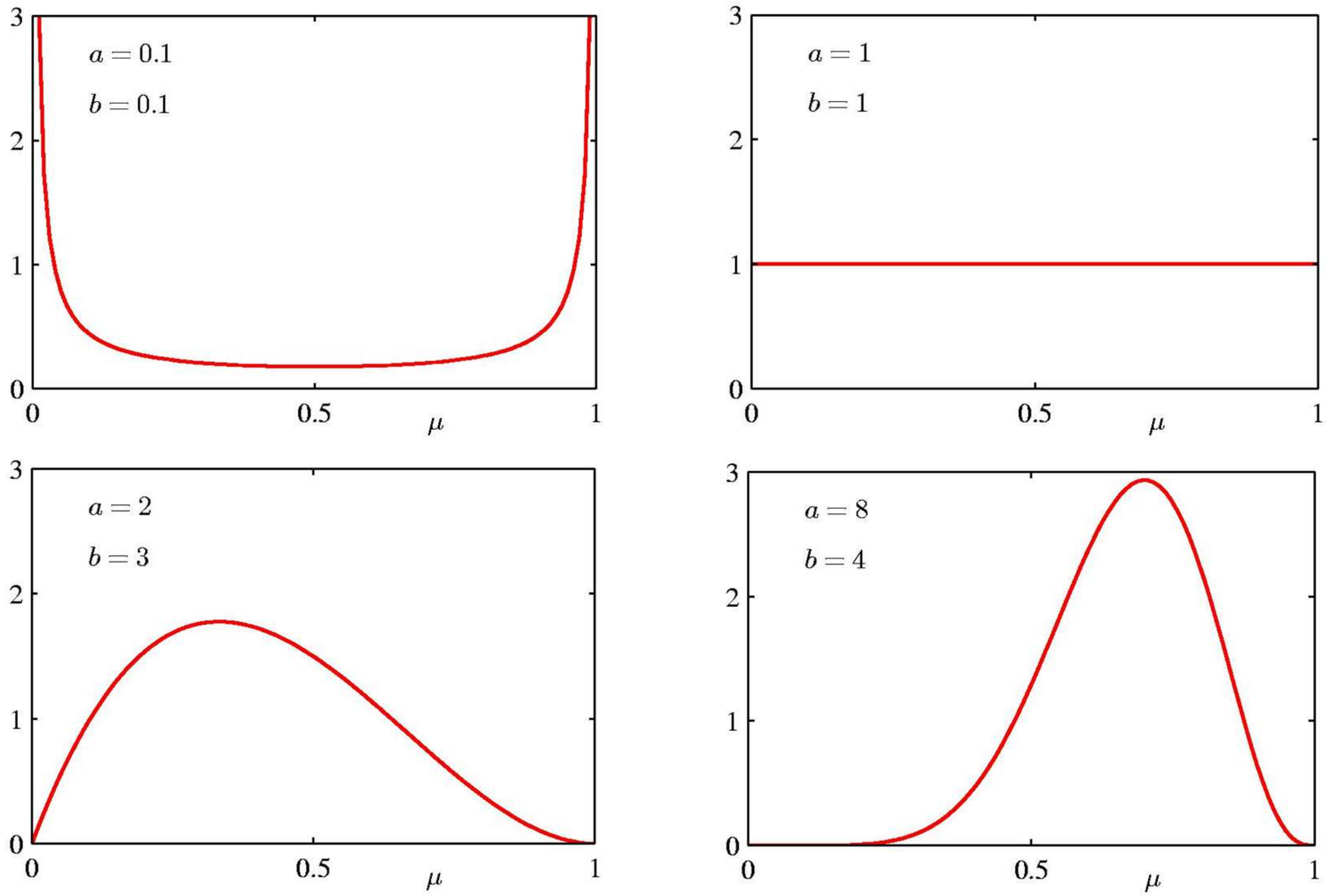
effective number of observations +1

The Beta distribution provides the *conjugate prior* for the Bernoulli distribution, i.e. the posterior distribution has the same form as the prior.

All distributions in the Exponential Family (includes multinomial, Gaussian, Poisson) have convenient conjugate priors (Bishop PRML 2.4).

Beta Distribution

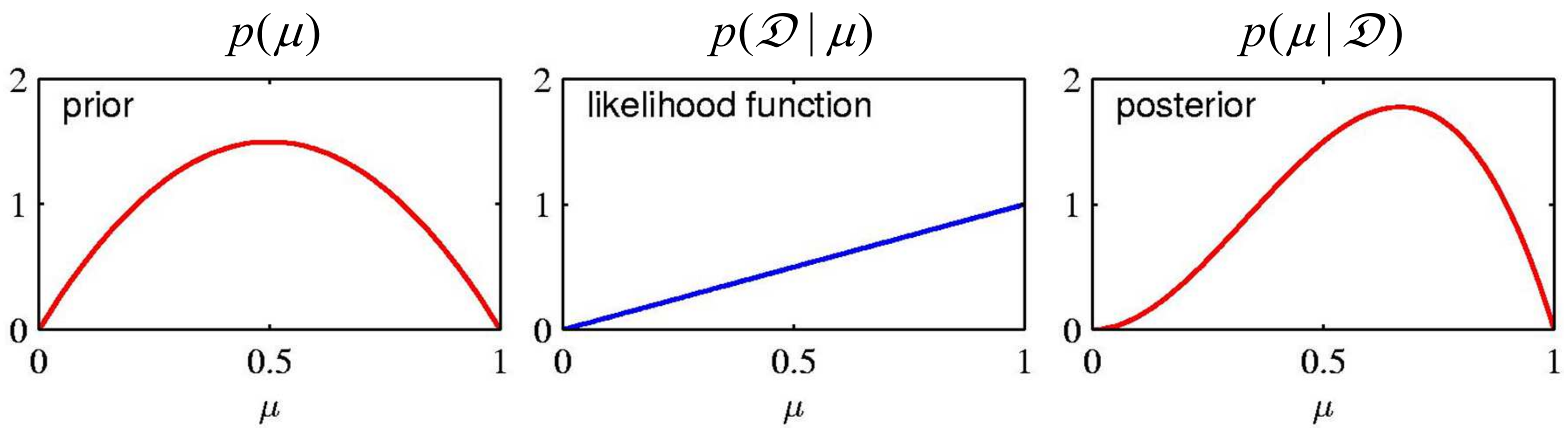
Which distribution is this?



Prior · Likelihood = Posterior

normalized

Example:



a=2, b=2 *Single observation H or x=1* *a=3, b=2*

Recall our earlier example of a Uniform prior, check this works…

Properties of the Posterior

As the size of the data set, N, grows

$$\mathbb{E}[\mu] = \frac{a_N}{a_N + b_N} = \frac{a_0 + m}{a_0 + m + b_0 + N - m} \to \frac{m}{N} = \mu_{ML}$$

$$\text{var}[\mu] = \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \to 0$$

This is typical behavior for Bayesian learning.

Prediction under the Posterior

What is the probability that the next coin toss will land heads up?

$$
\begin{aligned}
p(x=1|a_0,b_0,\mathcal{D}) &= \int_0^1 p(x=1|\mu)p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\
&= \int_0^1 \mu p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\
&= \mathbb{E}[\mu|a_0,b_0,\mathcal{D}] = \frac{a_N}{a_N+b_N}
\end{aligned}
$$

If we use our earlier example of a Uniform prior,
Observe $\mathcal{D} = (H,H,T,H) \Rightarrow a_N = 1+3,\ b_N = 1+1$

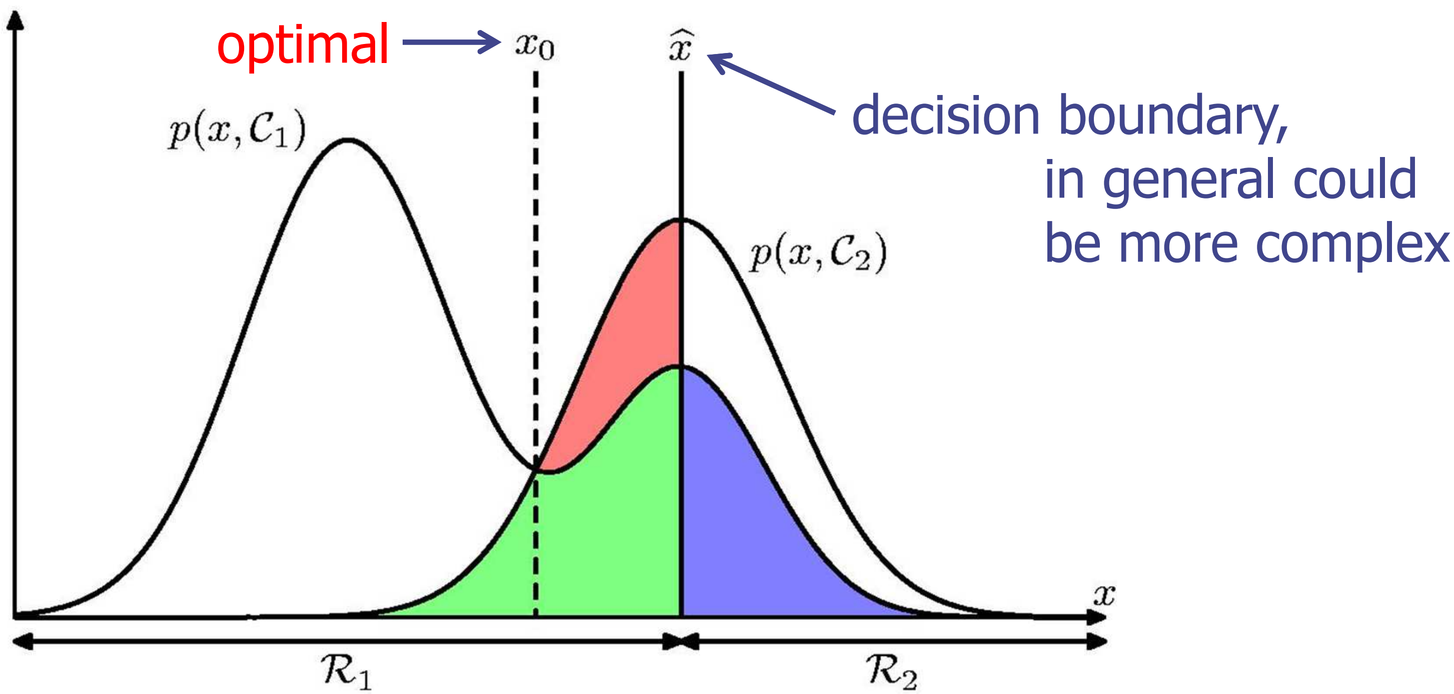
Now using posterior, P(next toss is a head) = $\frac{4}{4+2} = \frac{2}{3}$

Bayesian Decision Theory

- Initially assume just 2 classes $\mathcal{C}_1, \mathcal{C}_2$
- Given input data x we want to determine which class is optimal
- Various possible criteria
- Need $p(\mathcal{C}_k \mid x)$ either directly (discriminative)
- Or by Bayes, (generative) $p(\mathcal{C}_k \mid x) = \dfrac{p(x, \mathcal{C}_k)}{p(x)} = \dfrac{p(x \mid \mathcal{C}_k) p(\mathcal{C}_k)}{p(x)}$
- Divide input space into *decision regions* $\mathcal{R}_1, \mathcal{R}_2$ separated by *decision boundaries* such that $x \in \mathcal{R}_k \Rightarrow \text{assign } \mathcal{C}_k$
- How might we choose boundaries?

Criterion 1: Minimize Misclassification Rate

COMS4771, Columbia University
(Bishop PRML 1.5)



assign class 1 but should be 2 — assign class 2 but should be 1

$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}.
\end{aligned}
$$

$$p(x, \mathcal{C}_k) = p(\mathcal{C}_k \mid x)\, p(x)$$

Criterion 2: Minimize Expected Loss

Example loss matrix:

classify medical images as 'cancer' or 'normal'

Decision

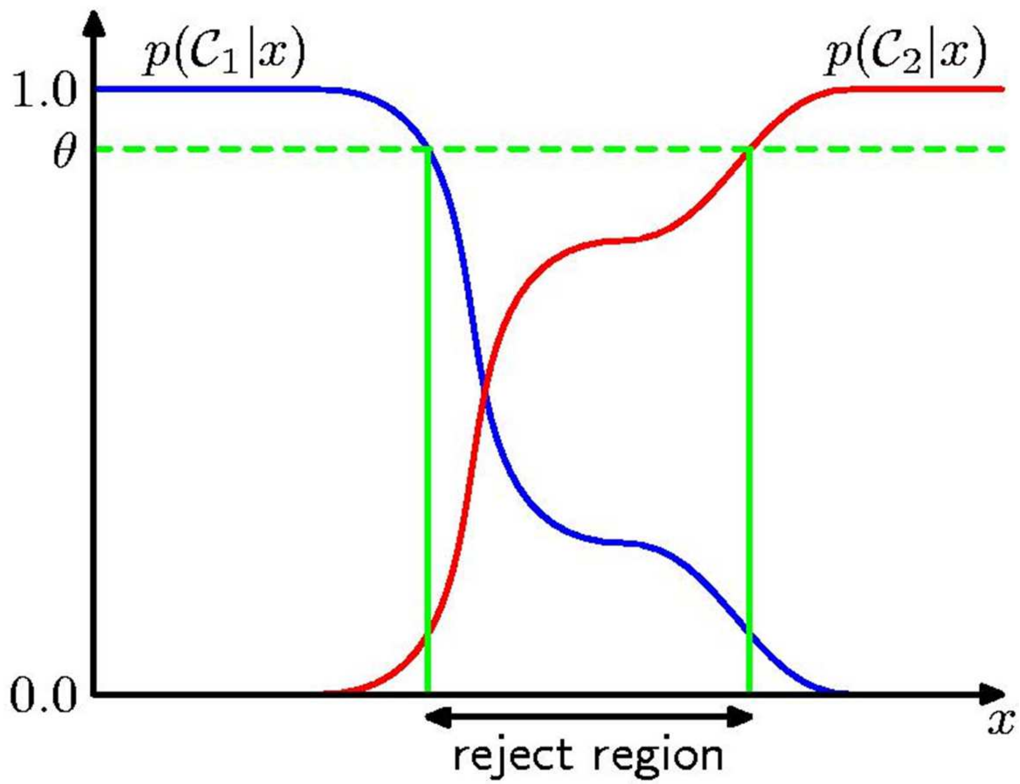
$$\text{Truth}\quad \begin{array}{c} \\ \text{cancer} \\ \text{normal} \end{array} \begin{array}{c} \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

Now $\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k)\,\mathrm{d}\mathbf{x}$

Choose regions $\mathcal{R}_j$ to minimize Expected Loss

$$\mathbb{E}[L] = \sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$$

Reject Option



If $p(\mathcal{C}_1 \mid x) \approx p(\mathcal{C}_2 \mid x)$ then less confident about assigning class.

One possibility is to reject or refuse to assign.