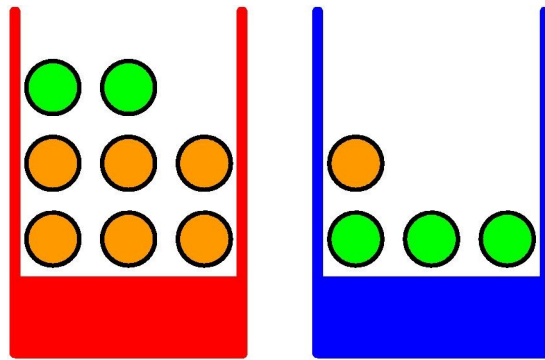


Crash course in probability theory and statistics – part 1



Machine Learning, Mon Apr 14, 2008

Motivation

Problem: To avoid relying on “magic” we need mathematics. For machine learning we need to quantify:

- Uncertainty in data measures and conclusions
- “Goodness” of model (when confronted with data)
- Expected error and expected success rates
- ...and many similar quantities...

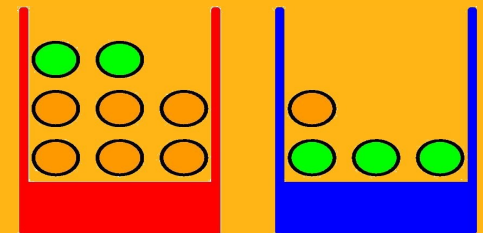
Motivation

Problem: To avoid relying on “magic” we need mathematics. For machine learning we need to quantify:

- Uncertainty in data measures and conclusions
- “Goodness” of model (when confronted with data)
- Expected error and expected success rates
- ...and many similar quantities...

Probability theory: Mathematical modeling when uncertainty or randomness is present.

$$P(X = x_i, Y = y_j) = p_{ij}$$

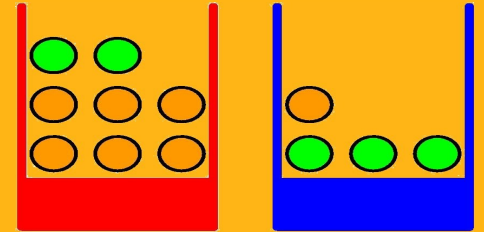


Motivation

Problem: To avoid relying on “magic” we need mathematics. For machine learning we need to quantify:

- Uncertainty in data
- “Goodness” of model
- Expected error and variance
- ...and many similar

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{n}$$



Probability theory: Mathematical modeling when uncertainty or randomness is present.

Statistics: The mathematics of collection of data, description of data, and inference from data

Introduction to probability theory

Notice: This will be an *informal* introduction to probability theory (measure theory out of scope for this course). No sigma-algebras, Borel-sets, etc.

For the purpose of this class, our intuition will be right ... in more complex settings it can be **very** wrong.

We leave the complex setups to the mathematicians and stick to “nice” models.

Introduction to probability theory

Notice: This will be an *informal* introduction to probability theory (measure theory out of scope for this course). No sigma-algebras, Borel-sets, etc.

For ... in This introduction will be based on **stochastic (random)** variables.

We leave the complex setups to the mathematicians and stick to “nice” models.

Introduction to probability theory

Notice: This will be an *informal* introduction to probability theory (measure theory out of scope for this course). No sigma-algebras, Borel-sets, etc.

For ... in This introduction will be based on **stochastic (random)** variables.

We and We ignore the underlying probability space (Ω, \mathcal{A}, p) .

Introduction to probability theory

Notice: This will be an *informal* introduction to probability theory (measure theory out of scope for this course). No sigma-algebras, Borel-sets, etc.

For ... in This introduction will be based on **stochastic (random)** variables.

We and We ignore the underlying probability space (Ω, \mathcal{A}, p) .

If X is the sum of two dice: $X(\omega) = D_1(\omega) + D_2(\omega)$



Introduction to probability theory

Notice: This will be an *informal* introduction to probability theory (measure theory out of scope for this course). No sigma-algebras, Borel-sets, etc.

For ... in This introduction will be based on **stochastic (random)** variables.

We and We ignore the underlying probability space (Ω, \mathcal{A}, p) .

If X is the sum of two dice: $X(\omega) = D_1(\omega) + D_2(\omega)$



We ignore the dice and only consider the variables – X , D_1 , and D_2 – and the values they take.

Discrete random variables

A **discrete random variable**, X , is a variable that can take values in a discrete (countable) set $\{x_i\}$.

The **probability** of X taking the value x_i is denoted $p(X=x_i)$ and satisfies $p(X=x_i) \geq 0$ for all i , $\sum_i p(X=x_i) = 1$, and for any subset $\{x_j\} \subseteq \{x_i\}$: $p(X \in \{x_j\}) = \sum_j p(x_j)$.

Discrete random variables

A **discrete random variable**, X , is a variable that can take values in a discrete (countable) set $\{x_i\}$.

The **probability** of X taking the value x_i is denoted $p(X=x_i)$ and satisfies $p(X=x_i) \geq 0$ for all i , $\sum_i p(X=x_i) = 1$, and for any subset $\{x_j\} \subseteq \{x_i\}$: $p(X \in \{x_j\}) = \sum_j p(x_j)$.

Intuition/interpretation: If we repeat an experiment (sampling a value for X) n times, and denote by n_i the number of times we observe $X=x_i$, then $n_i/n \rightarrow p(X=x_i)$ as $n \rightarrow \infty$.

Discrete random variables

A **discrete random variable**, X , is a variable that can take values in a discrete (countable) set $\{x_i\}$.

The **probability** $p(X=x_i)$ and satisfies $\sum p(X=x_i) = 1$ for any subset $\{x_j\} \subseteq \{x_i\}$. This is the **intuition** not a definition! (Definitions based on this ends up going in circles). The definitions are pure abstract math. Any real-world usefulness is pure luck.

Intuition/interpretation: If we repeat an experiment (sampling a value for X) n times, and denote by n_i the number of times we observe $X=x_i$, then $n_i/n \rightarrow p(X=x_i)$ as $n \rightarrow \infty$.

Discrete random variables

A **discrete random variable**, X , is a variable that can take values in a discrete (countable) set $\{x_i\}$.

The **probability** $p(X=x_i)$ is the probability that X takes the value x_i . This is the **intuition** not a definition! (Definitions based on this ends up going in circles).

subset $\{x_i\}$. We often simplify the notation and use both $p(X)$ and $p(x_i)$ for $p(X=x_i)$, depending on context.

Intuition and $p(x_i)$ for $p(X=x_i)$, depending on context. If we observe $X=x_i$, then $n_i/n \rightarrow p(X=x_i)$ as $n \rightarrow \infty$.

we observe $X=x_i$, then $n_i/n \rightarrow p(X=x_i)$ as $n \rightarrow \infty$.

Joint probability

If a random variable, Z , is a vector, $Z=(X, Y)$, we can consider its components separately.

The probability $p(Z=z)$ where $z = (x,y)$ is the **joint probability** of $X=x$ and $Y=y$ written $p(X=x, Y=y)$ or $p(x,y)$.

When **clear from context**, we write just $p(X, Y)$ or $p(x,y)$ and the notation is symmetric: $p(X, Y) = p(Y, X)$ and $p(x,y) = p(y,x)$.

The probability of $X \in \{x_i\}$ and $Y \in \{y_j\}$ becomes $\sum_i \sum_j p(x_i, y_j)$.

Marginal probability

The probability of $X=x_i$ regardless of the value of Y then becomes $\sum_j p(x_i, y_j)$ and is denoted the ***marginal probability*** of X and is written just $p(x_i)$.

The sum rule:

$$p(X) = \sum_Y p(X, Y) \quad (1.10)$$

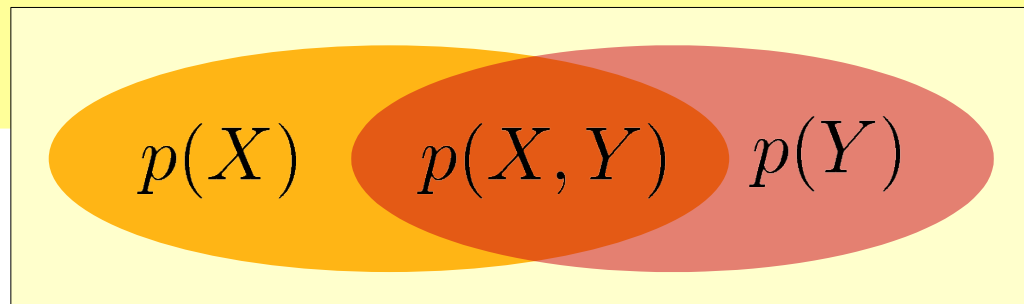
Conditional probability

The **conditional probability of X given Y** is written $P(X|Y)$ and is the quantity satisfying $p(X, Y) = p(X|Y)p(Y)$.

The product rule:

$$p(X, Y) = p(X|Y)P(Y) \quad (1.11)$$

When $p(Y) \neq 0$ we get $p(X|Y) = p(X, Y) / p(Y)$ with a simple interpretation.



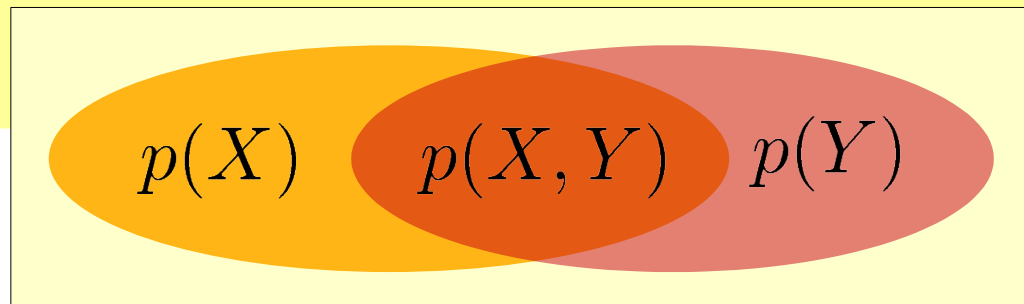
Conditional probability

The **conditional probability of X given Y** is written $P(X|Y)$ and is the quantity satisfying $p(X, Y) = p(X|Y)p(Y)$.

Intuition: Before we observe anything, the probability of X is $p(X)$ but after we observe Y it becomes $p(X|Y)$.

$$p(X, Y) = p(X|Y)P(Y) \quad (1.11)$$

When $p(Y) \neq 0$ we get $p(X|Y) = p(X, Y) / p(Y)$ with a simple interpretation.



Independence

When $p(X, Y) = p(X)p(Y)$ we say that X and Y are ***independent***.

In this case:

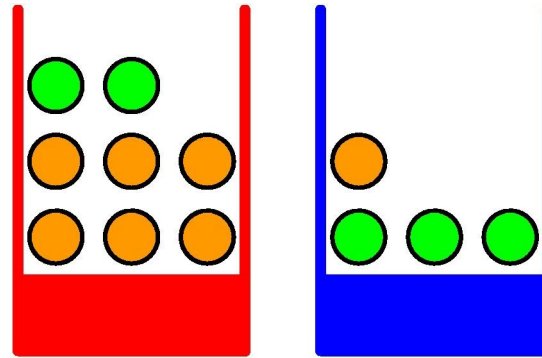
$$p(X|Y) = \frac{p(X, Y)}{p(Y)} = \frac{p(X)p(Y)}{p(Y)} = p(X)$$

Intuition/justification: Observing Y does not change the probability of X .

Example

B – colour of bucket

F – kind of fruit



$$p(B = r) = \frac{4}{10} \quad p(B = b) = \frac{6}{10}$$

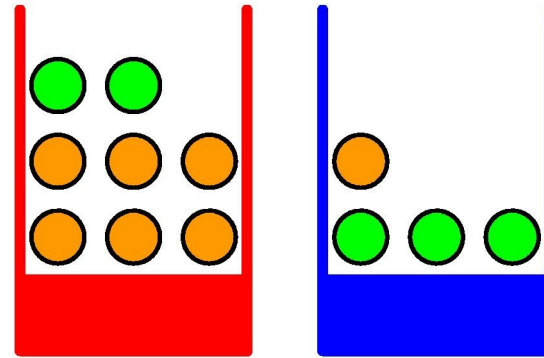
$$p(F = a|B = r) = \frac{2}{8} \quad p(F = a|B = b) = \frac{3}{4}$$

$$p(F = o|B = r) = \frac{6}{8} \quad p(F = o|B = b) = \frac{1}{4}$$

Example

B – colour of bucket

F – kind of fruit



$$p(B = r) = \frac{4}{10} \quad p(B = b) = \frac{6}{10}$$

$$p(F = a|B = r) = \frac{2}{8} \quad p(F = a|B = b) = \frac{3}{4}$$

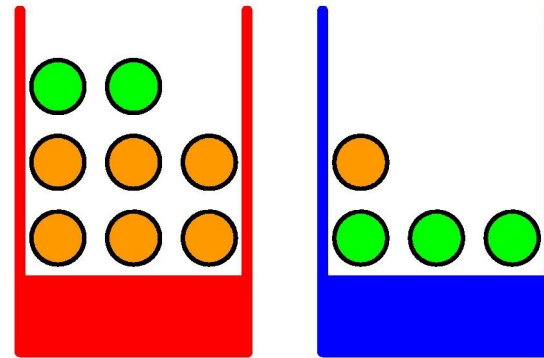
$$p(F = a, B = r) = p(F = a|B = r)p(B = r) = \frac{2}{8} \times \frac{4}{10} = \frac{1}{10} \quad = b) = \frac{1}{4}$$

$$p(F = a, B = b) = p(F = a|B = b)p(B = b) = \frac{2}{8} \times \frac{6}{10} = \frac{9}{20}$$

Example

B – colour of bucket

F – kind of fruit



$$p(B = r) = \frac{4}{10} \quad p(B = b) = \frac{6}{10}$$

$$p(F=a) = p(F=a, B=r) + p(F=a, B=b) = \frac{1}{10} + \frac{9}{10} = \frac{11}{20} \quad \frac{3}{4}$$

$$p(F=a, B=r) = p(F=a|B=r)p(B=r) = \frac{2}{8} \times \frac{4}{10} = \frac{1}{10} \quad = b) = \frac{1}{4}$$

$$p(F=a, B=b) = p(F=a|B=b)p(B=b) = \frac{2}{8} \times \frac{6}{10} = \frac{9}{20}$$

Bayes' theorem

Since $p(X, Y) = p(Y, X)$ (symmetry) and $p(X, Y) = p(Y|X)p(X)$ (product rule) it follows $p(Y|X)p(X) = p(X|Y)p(Y)$ or, when $p(X) \neq 0$:

Bayes' theorem:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \quad (1.12)$$

Sometimes written: $p(Y|X) \propto p(X|Y)p(Y)$ where $p(X) = \sum_Y p(X|Y)p(Y)$ is an implicit **normalising factor**.

Bayes' theorem

Since $p(X, Y) = p(Y, X)$ (symmetry) and $p(X, Y) = p(Y|X)p(X)$ (product rule) it follows $p(Y|X)p(X) = p(X|Y)p(Y)$ or when

Interpretation:

Prior to an experiment, the probability of Y is $p(Y)$

After observing X , the probability is $p(Y|X)$

Bayes' theorem tells us how to move from prior to posterior.

Sometimes written: $p(Y|X) \propto p(X|Y)p(Y)$ where

$p(X) = \sum_Y p(X|Y)p(Y)$ is an implicit **normalising factor**.

12)

Bayes' theorem

Since $p(X, Y) = p(Y, X)$ (symmetry) and $p(X, Y) = p(Y|X)p(X)$ (product rule) it follows $p(Y|X)p(X) = p(X|Y)p(Y)$ or when

Interpretation:

Prior to an experiment, the probability of Y is $p(Y)$

After o

This is possibly the most important equation in the entire class!

(12)

Bayes'

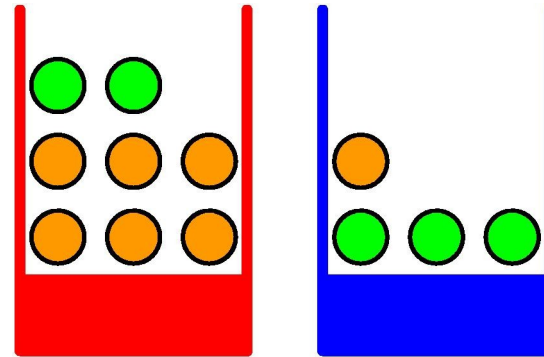
Sometimes written: $p(Y|X) \propto p(X|Y)p(Y)$ where

$p(X) = \sum_Y p(X|Y)p(Y)$ is an implicit **normalising factor**.

Example

B – colour of bucket

F – kind of fruit



If we draw an orange, what is the probability we drew it from the blue basket?

$$\begin{aligned} p(B = b|F = o) &= \frac{p(F = o|B = b)p(B = b)}{p(F = o)} \\ &= \frac{1/4 \times 6/10}{9/20} = \frac{1}{3} \end{aligned}$$

Continuous random variables

A **continuous random variable**, X , is a variable that can take values in \mathbb{R}^d .

The **probability density** of X is an integrable function $p(X)$ satisfying $p(x) \geq 0$ for all x and $\int p(x) dx = 1$.

The **probability** of $X \in S \subseteq \mathbb{R}^d$ is given by $p(S) = \int_S p(x) dx$.

Expectation

The **expectation** or **mean** of a function f of random variable X is a weighted average

$$\mathbb{E}[f] = \sum_x p(x) f(x) \qquad \mathbb{E}[f] = \int p(x) f(x) dx$$

For both discrete and continuous random variables:

$$\frac{1}{N} \sum_{n=1}^N f(x_n) \rightarrow \mathbb{E}[f] \qquad (1.35)$$

as $N \rightarrow \infty$ when $x_n \sim p(X)$.

Expectation

Intuition: If you repeatedly play a game with gain $f(x)$, your expected overall gain after n games will be $n\mathbb{E}[f]$.

The accuracy of this prediction increases with n .

It might not even be possible to “gain” $\mathbb{E}[f]$ in a single game.

Expectation

Intuition: If you repeatedly play a game with gain $f(x)$, your expected overall gain after n games will be $n\mathbb{E}[f]$.

The accuracy of this prediction increases with n .

It might not even be possible to “gain” $\mathbb{E}[f]$ in a single game.

Example: Game of dice with a fair dice, D value of dice, “gain” function $f(d) = d$.

$$\mathbb{E}[f] = \sum_{i=1}^6 \frac{i}{6} = 3.5$$

Variance

The **variance** of $f(x)$ is defined as $\mathbb{E} \left[(f(x) - \mathbb{E}[f(x)])^2 \right]$ and can be seen as a measure of variability around the mean.

The **covariance** of X and Y is defined as

$$\text{cov}[x, y] = \mathbb{E} [(x - \mathbb{E}[x]) (y - \mathbb{E}[y])]$$

and measures the variability of the two variables **together**.

Variance

The **variance** of $f(x)$ is defined as $\mathbb{E} \left[(f(x) - \mathbb{E}[f(x)])^2 \right]$ and can be seen as a measure of variability around the mean.

The **covariance** of X and Y is defined as

$$\text{cov}[x, y] = \mathbb{E} [(x - \mathbb{E}[x]) (y - \mathbb{E}[y])]$$

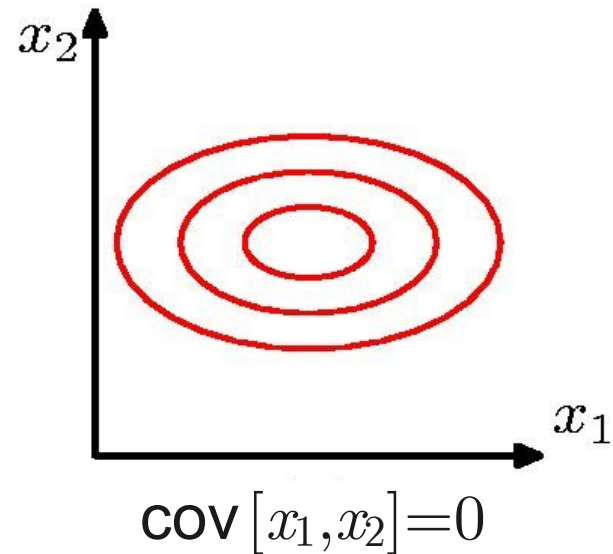
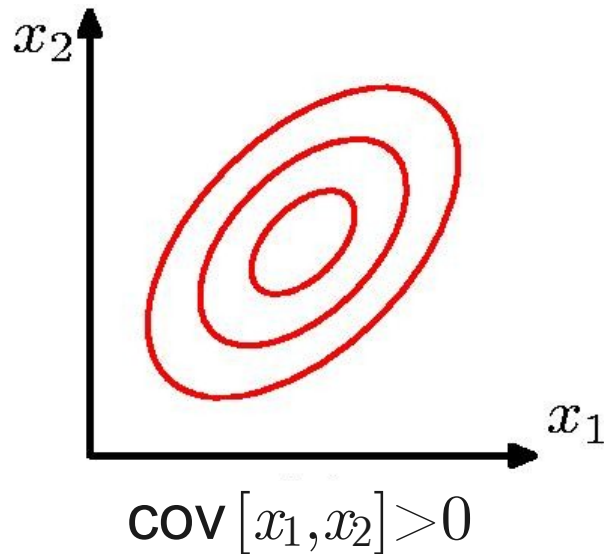
and measures the variability of the two variables **together**.

When $\text{cov}[x, y] > 0$, when x is above mean, y tends to be.

When $\text{cov}[x, y] < 0$, when x is above mean, y tends to be below.

When $\text{cov}[x, y] = 0$, x and y are **uncorrelated** (not necessarily independent; independence implies uncorrelated, though).

Covariance



When $\text{cov}[x, y] > 0$, when x is above mean, y tends to be.

When $\text{cov}[x, y] < 0$, when x is above mean, y tends to be below.

When $\text{cov}[x, y] = 0$, x and y are **uncorrelated** (not necessarily independent; independence implies uncorrelated, though).

Parameterized distributions

Many distributions are governed by a few parameters.

E.g. coin tossing (Bernoulli distribution) governed by the probability of “heads”.

Binomial distribution:
number of “heads” k out of n coin tosses:

$$p(k | n, \theta) \propto \theta^k (1 - \theta)^{n-k}$$



$$p(\text{head} | \theta) = \theta$$

$$p(\text{tail} | \theta) = 1 - \theta$$

Parameterized distributions

Many distributions are

We can think of a parameterized distribution as a conditional distribution.

The function $x \rightarrow p(x | \theta)$ is the **probability** of observation x given parameter θ .

The function $\theta \rightarrow p(x | \theta)$ is the **likelihood** of parameter θ given observation x . Sometimes written $\text{Lhd}(\theta | x) = p(x | \theta)$.

$$p(k | n, \theta) \propto \theta^k (1 - \theta)^{n-k}$$

θ

Parameterized distributions

Many distributions are

generated by a parameter θ . We can think of a parameterized distribution as a conditional distribution.

The function $x \rightarrow p(x | \theta)$ is the **probability** of observation x given parameter θ .

The function $\theta \rightarrow p(x | \theta)$ is the **likelihood** of parameter θ given observation x . Sometimes written $\text{Lhd}(\theta | x) = p(x | \theta)$.

The likelihood, in general, is not a probability distribution.

θ

Parameter estimation

Generally, parameters are not known but must be **estimated** from observed data.

Maximum Likelihood (ML):

$$\hat{\theta} = \underset{\theta}{\operatorname{maxarg}} p(x | \theta)$$

Maximum A Posteriori (MAP):

(A Bayesian approach assuming a distribution over parameters).

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{maxarg}} p(\theta | x) \\ &= \underset{\theta}{\operatorname{maxarg}} p(x | \theta)p(\theta)\end{aligned}$$

Fully Bayesian:

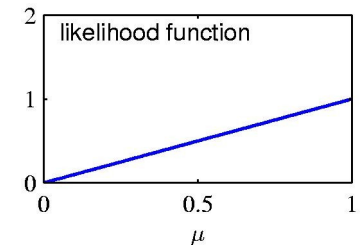
(Estimates a distribution rather than a parameter).

$$\begin{aligned}\hat{p}(y | x) &= \int p(y | \theta)p(\theta | x) d\theta \\ &\propto \int p(y | \theta)p(x | \theta)p(\theta) d\theta\end{aligned}$$

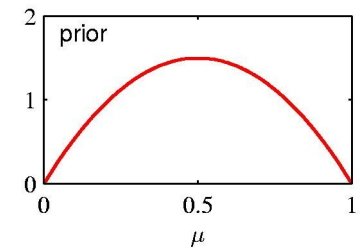
Parameter estimation

Example: We toss a coin and get a “head”. Our model is a binomial distribution; x is one “head” and θ the probability of a “head”.

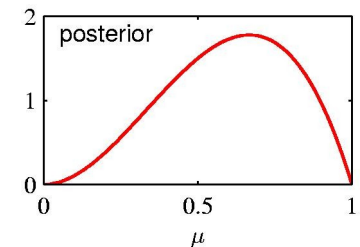
Likelihood: $\text{lh}(\theta | x) = p(x | \theta) \propto \theta^1 (1 - \theta)^0$



Prior: $p(\theta) \propto \theta^\alpha (1 - \theta)^\beta$



Posterior: $p(\theta | x) \propto \theta^{\alpha+1} (1 - \theta)^\beta$



Parameter estimation

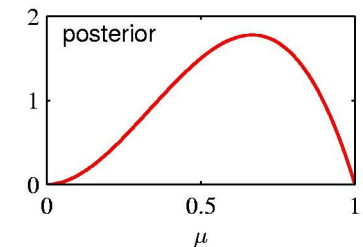
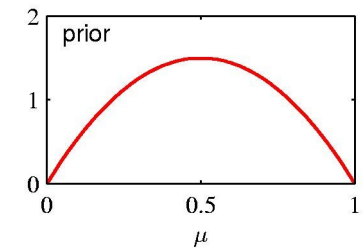
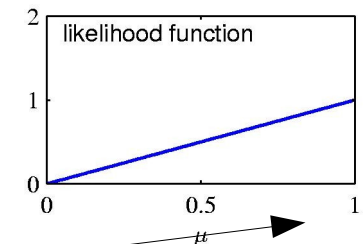
Example: We toss a coin and get a “head”. Our model is a binomial distribution; x is one “head” and θ the probability of a “head”.

Likelihood: $\text{lh}(\theta | x) = p(x | \theta) \propto \theta^1 (1 - \theta)^0$

ML estimate

Prior: $p(\theta) \propto \theta^\alpha (1 - \theta)^\beta$

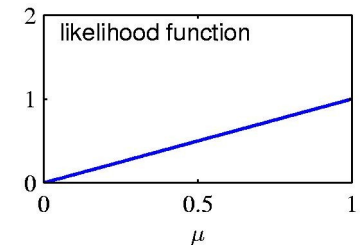
Posterior: $p(\theta | x) \propto \theta^{\alpha+1} (1 - \theta)^\beta$



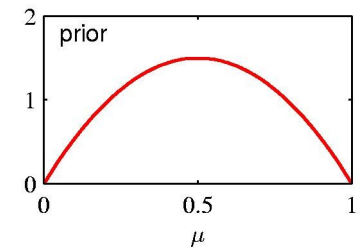
Parameter estimation

Example: We toss a coin and get a “head”. Our model is a binomial distribution; x is one “head” and θ the probability of a “head”.

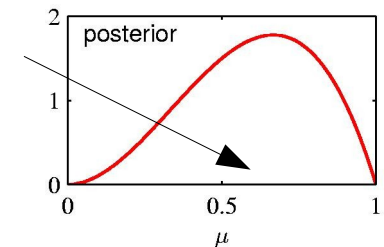
Likelihood: $\text{lh}(\theta | x) = p(x | \theta) \propto \theta^1 (1 - \theta)^0$



Prior: $p(\theta) \propto \theta^\alpha (1 - \theta)^\beta$



Posterior: $p(\theta | x) \propto \theta^{\alpha+1} (1 - \theta)^\beta$ MAP estimate



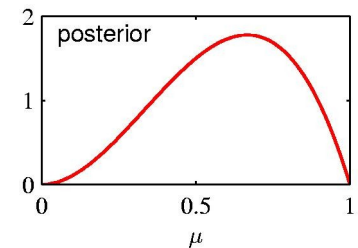
Parameter estimation

Example: We toss a coin and get a “head”. Our model is a binomial distribution; x is one “head” and θ the probability of a “head”.

Fully Bayesian approach:

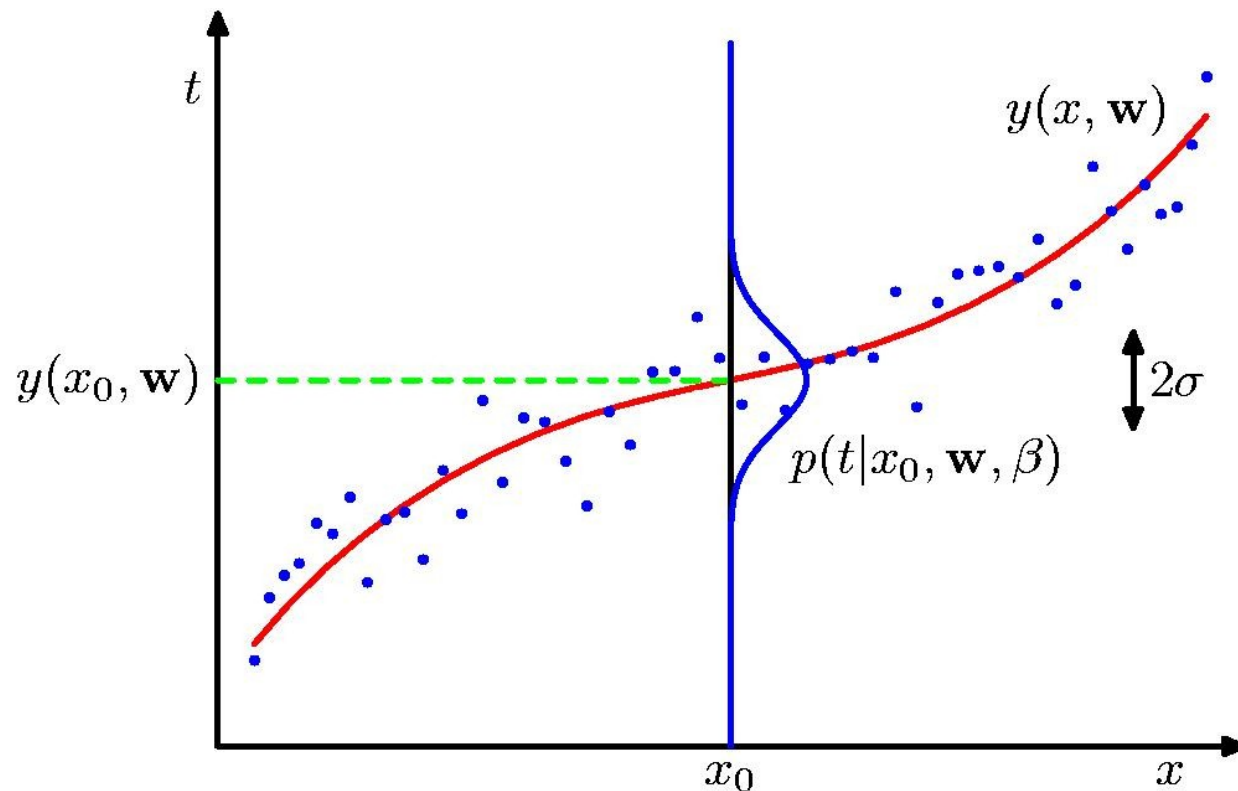
$$\begin{aligned}\hat{p}(y | x) &= \int_0^1 p(y | x)p(\theta | x) d\theta \\ &\propto \int_0^1 \theta^{\alpha+1+\mathbf{1}_{y=h}} (1 - \theta)^{\beta+1-\mathbf{1}_{y=h}} d\theta \\ &= B(\alpha + 1 + \mathbf{1}_{y=h}, \beta + 1 - \mathbf{1}_{y=h})\end{aligned}$$

Posterior: $p(\theta | x) \propto \theta^{\alpha+1} (1 - \theta)^{\beta}$



Predictions

Assume now known joint distribution $p(x, t | \theta)$ of **explanatory** variable x and **target** variable t . When observing new x we can use $p(t | x, \theta)$ to make predictions about t .



Decision theory

Based on $p(x, t | \theta)$ we often need to make decisions.

This often means taking one of a small set of actions $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_k$ based on observed x .

Assume that the target variable is in this set, then we make decisions based on $p(t | x, \theta) = p(\mathcal{A}_i | x, \theta)$.

Put in a different way: we use $p(x, t | \theta)$ to classify x into one of k classes, \mathcal{C}_i .

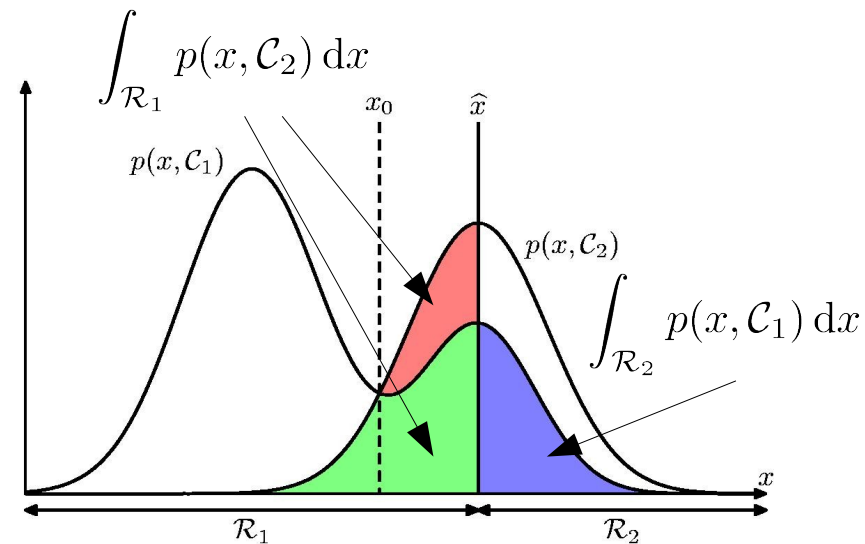
Decision theory

We can approach this by splitting the input into regions, \mathcal{R}_i , and make decisions based on these:

In \mathcal{R}_1 go for \mathcal{C}_1 ; in \mathcal{R}_2 go for \mathcal{C}_2 .

Choose regions to minimize classification errors:

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



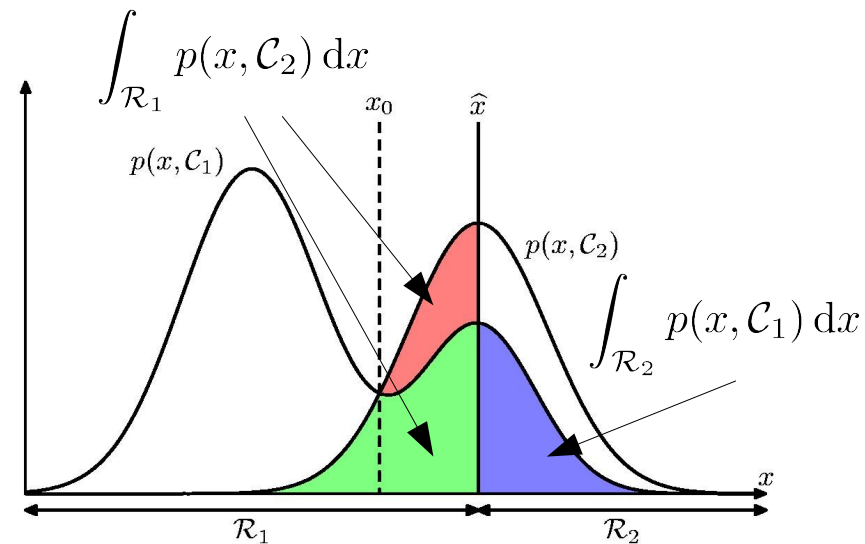
Decision theory

We can approach this by splitting the input into regions, \mathcal{R}_i , and make decisions based on these:

In \mathcal{R}_1 go for \mathcal{C}_1 ; in \mathcal{R}_2 go for \mathcal{C}_2 .

Choose regions to minimize classification errors:

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



Red and green mis-classifies \mathcal{C}_2 as \mathcal{C}_1

Blue mis-classifies \mathcal{C}_1 as \mathcal{C}_2

At x_0 red is gone and $p(\text{mistake})$ is minimized

Decision theory

We can approach this by splitting the input into regions, \mathcal{R}_i , and make decisions based on these:

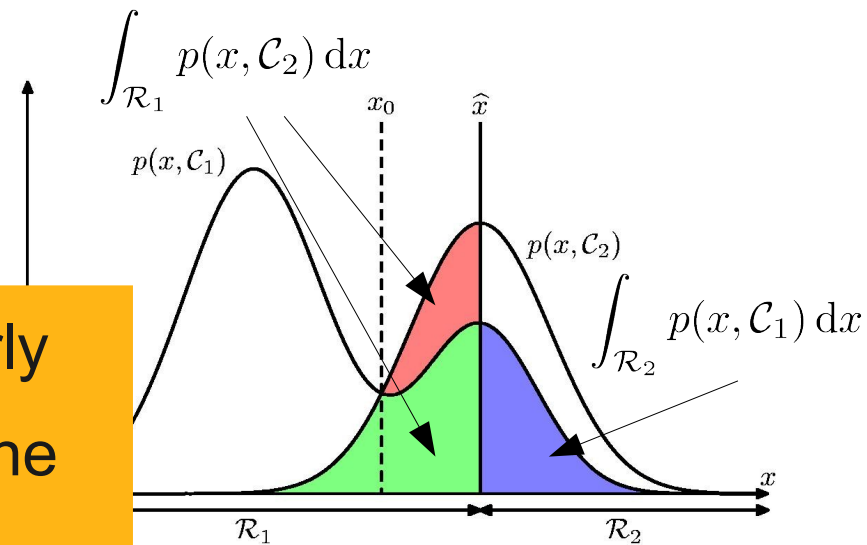
In \mathcal{R}_1 go for \mathcal{C}_1 ; in \mathcal{R}_2 go for \mathcal{C}_2 .

Choose regions to minimize classification errors:

x_0 is where $p(x, \mathcal{C}_1) = p(x, \mathcal{C}_2)$ or similarly $p(\mathcal{C}_1 | x)p(x) = p(\mathcal{C}_2 | x)p(x)$ so we get the intuitive pleasing:

$$\mathcal{R}_1 = \{x \mid p(\mathcal{C}_1 | x) > p(\mathcal{C}_2 | x)\}$$

$$\mathcal{R}_2 = \{x \mid p(\mathcal{C}_2 | x) > p(\mathcal{C}_1 | x)\}$$



Model selection

Where do we get $p(t, x | \theta)$ from in the first place?

Model selection

Where do we get $p(t, x | \theta)$ from in the first place?

There is no ***right*** model – a fair coin or fair dice is as unrealistic as a spherical cow!

Model selection

Where do we get $p(t, x | \theta)$ from in the first place?

There is no ***right*** model – a fair coin or fair dice is as unrealistic as a spherical cow!

Sometimes there are obvious candidates to try – either for the joint or conditional probabilities $p(x, t | \theta)$ or $p(t | x, \theta)$.

Sometimes we can try a "generic" model – linear models, neural networks, ...

Model selection

Where do we get $p(t, x | \theta)$ from in the first place?

There is no ***right*** model – a fair coin or fair dice is as unrealistic as a spherical cow!

Sometimes there are obvious candidates to try – either for the joint or conditional probabilities $p(x, t | \theta)$ or $p(t | x, \theta)$.

Sometimes we can try a "generic" model – linear models, neural networks, ...

This is the topic of most of this class!

Model selection

Where do we get $p(t, x | \theta)$ from in the first place?

There is no ***right*** model – a fair coin or fair dice is as unrealistic as a spherical cow!

Model selection

Where do we get $p(t, x | \theta)$ from in the first place?

There is no *right* model – a fair coin or fair dice is as unrealistic as a spherical cow!

But some models are more *useful* than others.

Model selection

Where do we get $p(t, x | \theta)$ from in the first place?

There is no ***right*** model – a fair coin or fair dice is as unrealistic as a spherical cow!

But some models are more ***useful*** than others.

If we have several models, how do we measure the usefulness of each?

Model selection

Where do we get $p(t, x | \theta)$ from in the first place?

There is no ***right*** model – a fair coin or fair dice is as unrealistic as a spherical cow!

But some models are more ***useful*** than others.

If we have several models, how do we measure the usefulness of each?

A good measure is prediction accuracy on new data.

Model selection

If we compare two models, we can take a maximum likelihood approach:

$$M = \operatorname{argmax}_M p(t, x \mid M)$$

or a Bayesian approach:

$$M = \operatorname{argmax}_M p(t, x \mid M)p(M)$$

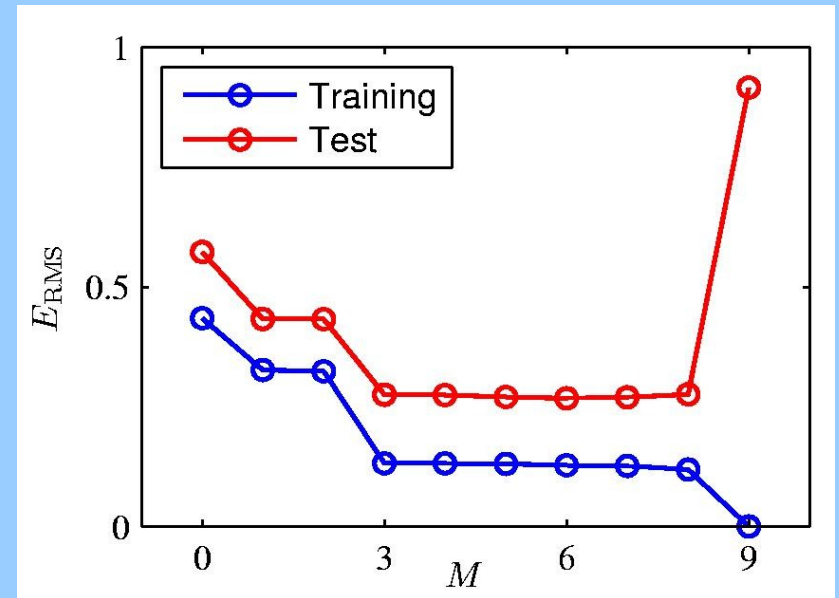
just as for parameters.

Model selection

If we compare two models, we can take a maximum likelihood approach:

But there is an *over fitting* problem:

Complex models often fit training data better *without generalizing better!*

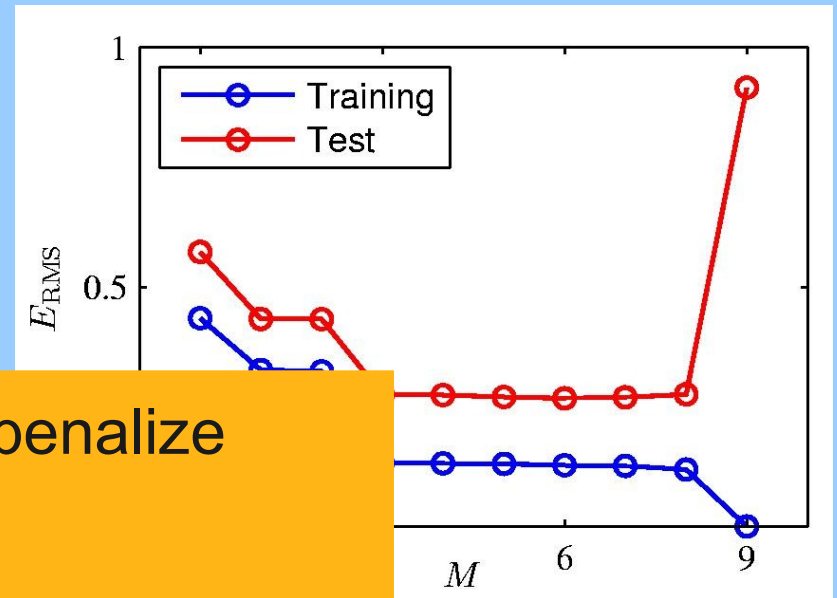


Model selection

If we compare two models, we can take a maximum likelihood approach:

But there is an **over fitting** problem:

Complex models often fit training data better **without generalizing better!**



In Bayesian approach, use $p(M)$ to penalize complex models

In ML approach, use some **Information Criteria** and maximize $\ln p(t, x | M) - \text{penalty}(M)$.

Model selection

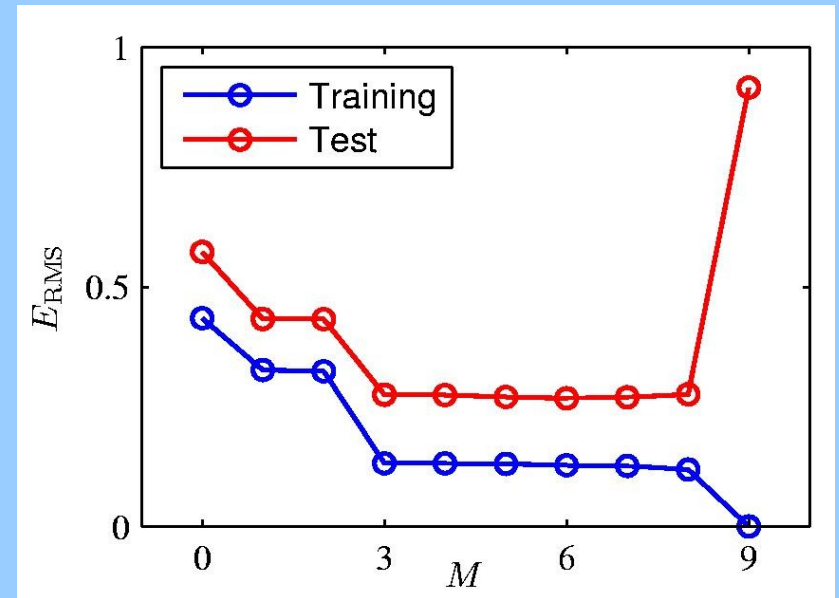
If we compare two models, we can take a maximum likelihood approach:

But there is an *over fitting* problem:

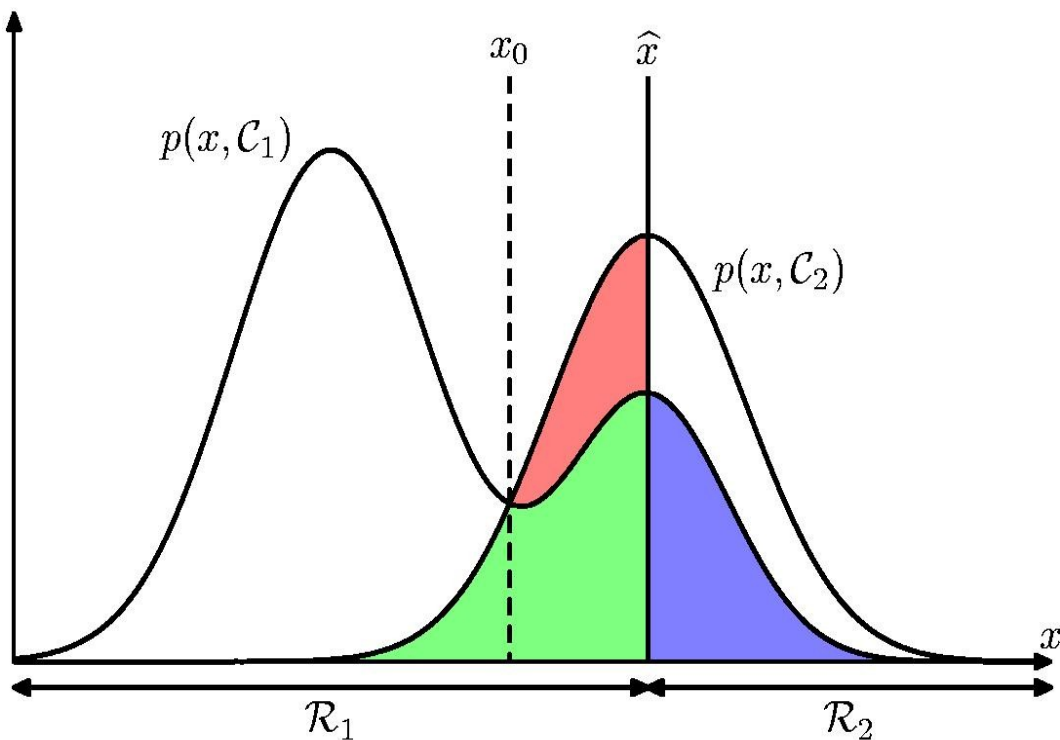
Complex models often fit training

Or more empirical approach: Use some method of splitting data into *training data* and *test data* and pick model that performs best on test data.

(and retrain that model with the full dataset).



Summary



- Probabilities
- Stochastic variables
- Marginal and conditional probabilities
- Bayes' theorem
- Expectation, variance and covariance
- Estimation
- Decision theory and model selection