Inference: Decoding & Regression

Banyuls school on « Advanced Computational Analysis for Behavioral and Neurophysiological Recordings »

B. Bathellier and R. Monasson

1. A primer on probabilities/conditional probabilities
2. Bayesian inference
3. Illustrations:
    - Laplace birth rate problem
    - Decoding of position from hippocampal place cells
4. Regression and logistic regression
5. Priors and cross-validation

Probabilities and conditional probabilities

Dice : faces x = 1,2,3,4,5,6 (here, x = 3)

Notation: Probability (face = x) $\equiv p(x)$

Properties : $p(x) \geq 0, \sum_{x=1,\ldots,6} p(x) = 1$

Unbiased dice : each face is equally likely through a draw

$$p(1) = p(2) = \ldots = p(6) = \frac{1}{6}$$

Probabilities and conditional probabilities

Two dices : faces x_1 and x_2 (here, $x_1 = x_2 = 6$)

Joint probability (1st face = x_1 & 2nd face = x_2) ≡ $p(x_1, x_2)$

Two important considerations:

- Care about one event only -> definition of marginal probability

$$p_1(x_1) \equiv \sum_{x_2} p(x_1, x_2), \; p_2(x_2) \equiv \sum_{x_1} p(x_1, x_2)$$

- Ask whether events are independent, i.e. are the two faces correlated or not?

$$p(x_1, x_2) = p_1(x_1) \times p_2(x_2) \; ?$$

An example of dependent events

Box *A*

20 plain cookies
+
20 chocolate cookies

Box *B*

10 plain cookies
+
30 chocolate cookies

One picks up uniformly at random one box and one cookie out of this box

Box*:* x_1 = *A or B* $\rightarrow$ $p_1(A) = p_1(B) = \frac{1}{2}$

Cookie type: x_2 = *plain* or *chocolate* $\rightarrow$ $p_2(plain) = \frac{3}{8}$

$p_2(chocolate) = \frac{5}{8}$

An example of dependent events

Box *A*

20 plain cookies
+
20 chocolate cookies

Box *B*

10 plain cookies
+
30 chocolate cookies

One picks up uniformly at random one box and one cookie out of this box

Box: $x_1 = A$ or $B \quad \rightarrow \quad p_1(A) = p_1(B) = \frac{1}{2}$

Cookie type: $x_2 = plain$ or $chocolate \quad \rightarrow \quad p_2(plain) = \frac{3}{8}$

$$p_2(chocolate) = \frac{5}{8}$$

but $p_2(chocolate|A) = \frac{1}{2} < p_2(chocolate) = \frac{5}{8} < p_2(chocolate|B) = \frac{3}{4}$

Characterization of dependent events

Definition of conditional probability: $$p_2\left(x_2 \middle| x_1\right) \equiv \frac{p(x_1, x_2)}{p_1(x_1)}$$

Exercise 1: *check that conditional probability is normalized!*

Characterization of dependent events

Definition of conditional probability: $$p_2\left(x_2 \middle| x_1\right) \equiv \frac{p(x_1, x_2)}{p_1(x_1)}$$

Exercise 1: *check that conditional probability is normalized!*

Box A

20 plain cookies
+
20 chocolate cookies

$p_2(x_2 = \text{plain} \mid x_1 = A) = 1/2,$
$p_2(x_2 = \text{chocolate} \mid x_1 = A) = 1/2$

Box B

10 plain cookies
+
30 chocolate cookies

$p_2(x_2 = \text{plain} \mid x_1 = B) = 1/4,$
$p_2(x_2 = \text{chocolate} \mid x_1 = B) = 3/4$

A side remark: mutual information

The dependence between events (or event distributions) is characterized through the **mutual information**

$$MI(1,2) = \sum_{x_1 x_2} p(x_1, x_2) \log_2 \left[\frac{p(x_1, x_2)}{p_1(x_1) p_2(x_2)} \right]$$

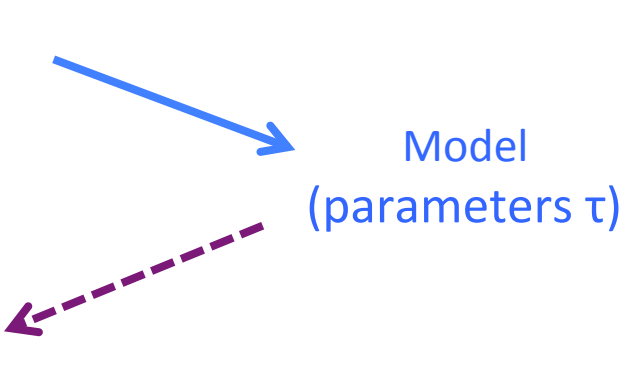
in bits

This quantity measures how much information one has on one variable from knowledge of the other one.

(1) Always positive (or null)
(2) Degraded through processing: $x_2 \rightarrow x_3$ then $MI(1,3) \leq MI(1,2)$

Exercise 2: *compute MI(box,cookie type) in bits*

Bayesian inference



Probabilistic description: joint distribution of σ & τ

A priori distribution of the model parameters

$$p(\sigma,\tau) = p(\sigma|\tau) \times p(\tau)$$

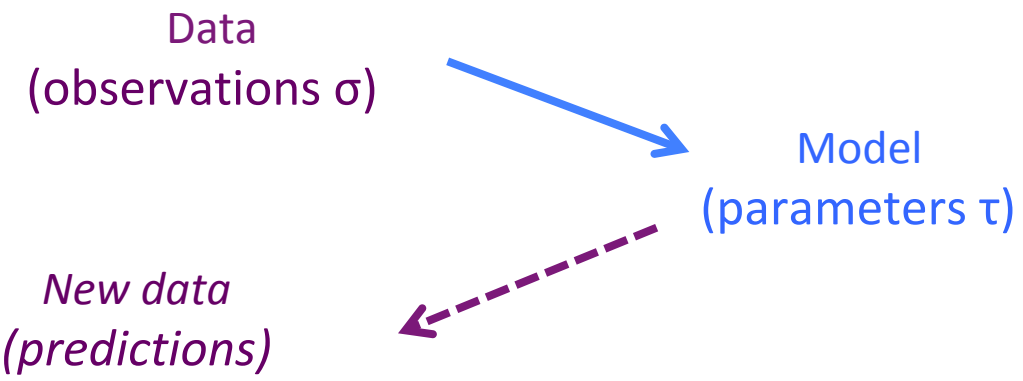
Likelihood of model parameters

Probability that data can be generated by model τ

$$= p(\tau|\sigma) \times p(\sigma)$$

A posteriori distribution of the model parameters
(can be sampled, maximized, …)

Bayesian inference



Probabilistic description: joint distribution of σ & τ

A priori distribution of the model parameters

$$p(\sigma,\tau) = p(\sigma|\tau) \times p(\tau)$$

Likelihood of model parameters

$$= p(\tau|\sigma) \times p(\sigma)$$

Probability that data can be generated by model τ

A posteriori distribution of the model parameters
(can be sampled, maximized, …)

Bayesian inference formula:

$$p(\tau|\sigma) = \frac{p(\sigma|\tau) \times p(\tau)}{p(\sigma)}$$

An illustration

Application to the chocolate/plain cookie problem:

$$P(A \mid \text{Chocolate}) = \frac{P(\text{Chocolate} \mid A) \times P_{\text{prior}}(A)}{P(\text{Chocolate})} = \frac{1/2 \times 1/2}{5/8} = \frac{2}{5}$$

$$P(B \mid \text{Chocolate}) = \frac{P(\text{Chocolate} \mid B) \times P_{\text{prior}}(B)}{P(\text{Chocolate})} = \frac{3/4 \times 1/2}{5/8} = \frac{3}{5}$$

Laplace and the birth rate of boys & girls

Historical example: « proof » by Laplace that the female and male birth rates are different

Data: Nbs of girls born in Paris from 1745 to 1770 : 245,945

… boys … : 251,527

σ = nb. of female births, n = total number of births

Laplace and the birth rate of boys & girls

Historical example: « proof » by Laplace that the female and male birth rates are different

Data: Nbs of girls born in Paris from 1745 to 1770 : 245,945
… boys … : 251,527

σ = nb. of female births, n = total number of births

Inference: τ = probability that a newborn baby is a girl

- Prior distribution: uniform over τ in [0;1]

Laplace and the birth rate of boys & girls

Historical example: « proof » by Laplace that the female and male birth rates are different

Data: Nbs of girls born in Paris from 1745 to 1770 : 245,945
… boys … : 251,527

σ = nb. of female births, n = total number of births

Inference: τ = probability that a newborn baby is a girl

- Prior distribution: uniform over τ in [0;1]
- Likelihood: $p(\sigma|\tau) = \binom{n}{\sigma} \tau^{\sigma} (1-\tau)^{n-\sigma}$

Laplace and the birth rate of boys & girls

Historical example: « proof » by Laplace that the female and male birth rates are different

Data: Nbs of girls born in Paris from 1745 to 1770 : 245,945
… boys … : 251,527

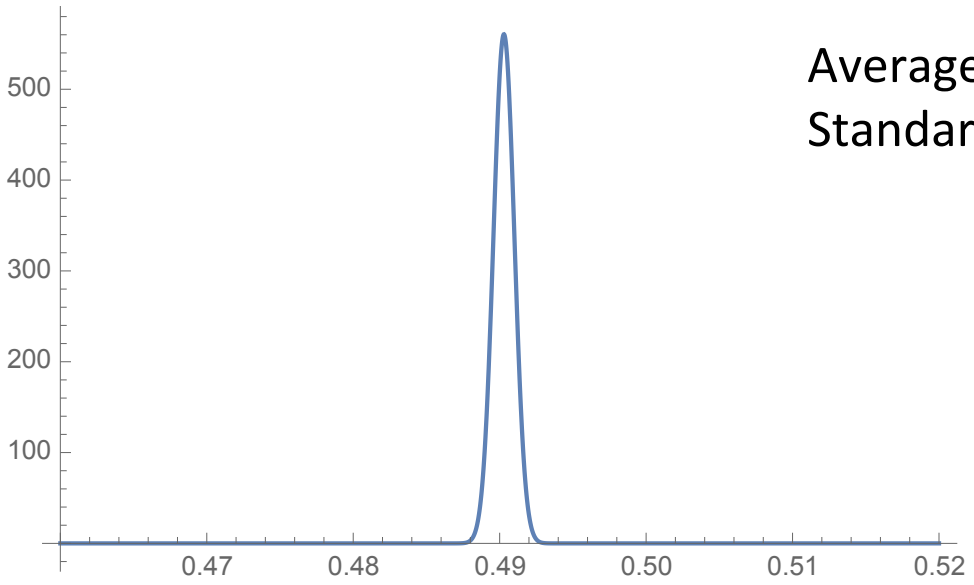
σ = nb. of female births, n = total number of births

Inference: τ = probability that a newborn baby is a girl

- Prior distribution: uniform over τ in [0;1]
- Likelihood: $p(\sigma|\tau) = \binom{n}{\sigma} \tau^{\sigma} (1-\tau)^{n-\sigma}$
- Bayes: $p(\tau|\sigma) = \frac{\tau^{\sigma}(1-\tau)^{n-\sigma}}{p(\sigma)}$

Laplace and the birth rate of boys & girls

Posterior distribution:

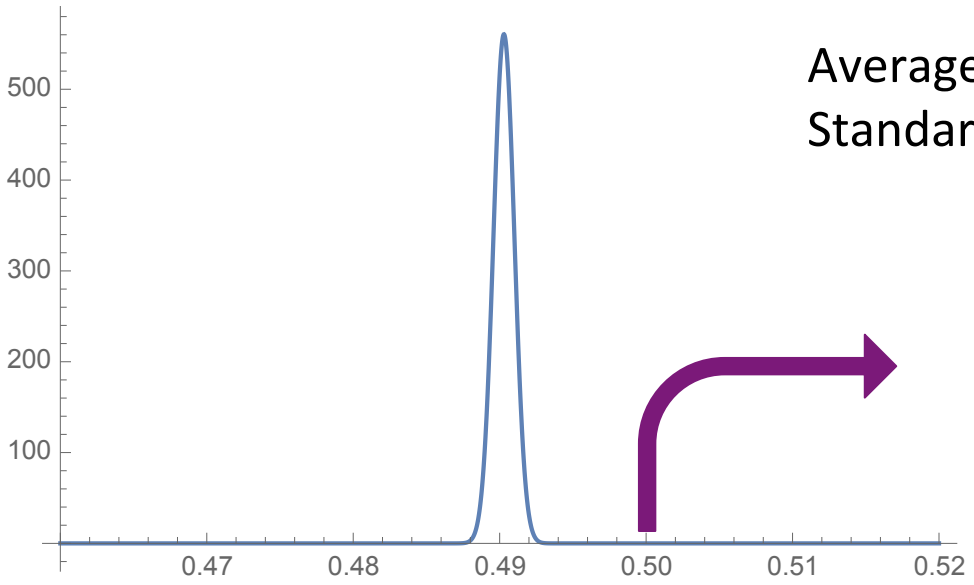


Average $\tau = 0.490291$

Standard deviation $\tau = 0.007117$

Laplace and the birth rate of boys & girls

Posterior distribution:



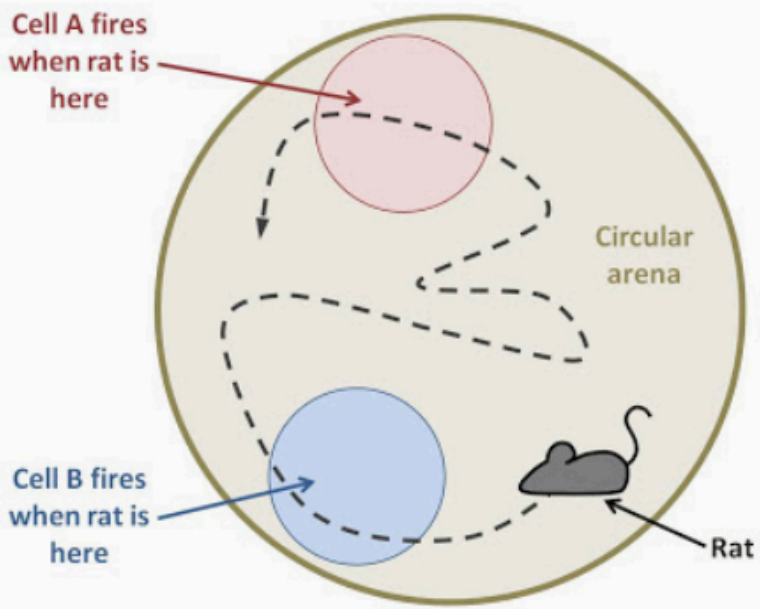
Average τ = 0.490291
Standard deviation τ = 0.007117

Probability that τ exceeds 0.5 = $\int_{0.5}^{1} d\tau\, p(\tau|\sigma) \approx 10^{-42}$

Extremely unlikely!

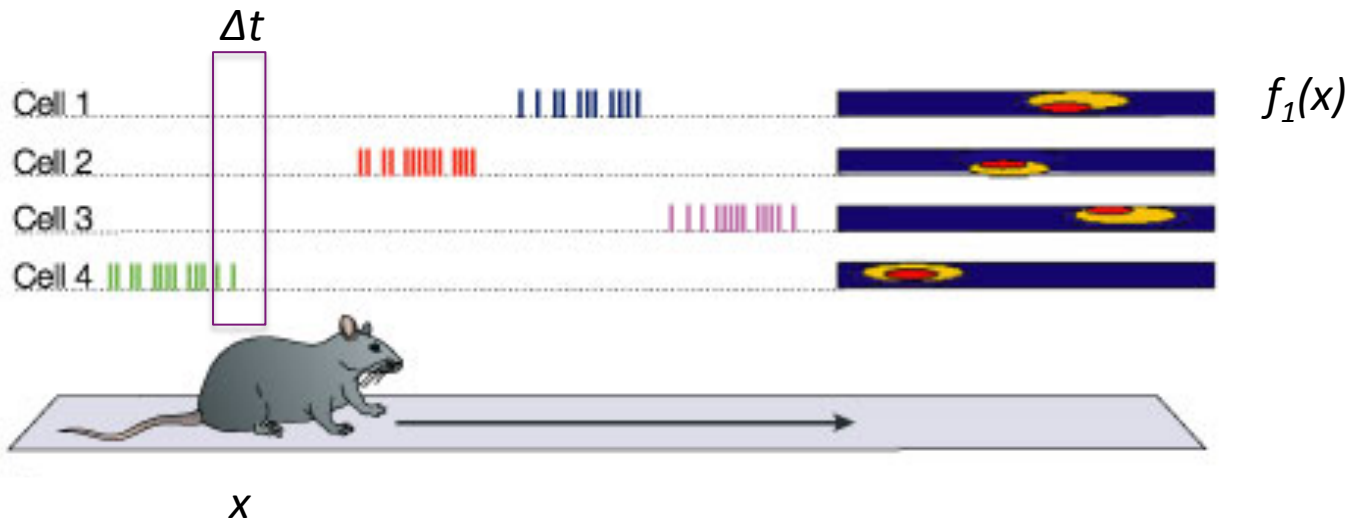
Decoding of position from HPC activity

Place cells in the hippocampus regions CA1 and CA3 present spatially-located firing fields.



- Place fields are retrieved when the animal is placed in the same environment after days
- Stable in dark and against limited changes of environment
- Low-dimensional projections of context-dependent place fields in complex, high-D space

Decoding of position from HPC activity



- Assume place cells are Poisson (?): $P_i\left(s_{i,t}\middle|x\right) = e^{-f_i(x)\Delta t}\dfrac{\left(f_i(x)\Delta t\right)^{s_{i,t}}}{s_{i,t}!}$

- Assume place cell activities are independent (conditional to position x):

$$P\left(\{s_{i,t}\}\middle|x\right) = \prod_i P_i\left(s_{i,t}\middle|x\right)$$

Decoding of position from HPC activity

- Prior over trajectory: $P_{prior}(x_t) =$ uniform over the environment

- Find most likely position:

MAP=ML decoding

$$x_t = \underset{x}{\operatorname{argmax}} \left[P\left(\{s_{i,t}\} \middle| x \right) \times P_{prior}(x_t) \right]$$

Decoding of position from HPC activity

- Prior over trajectory: $P_{prior}(x_t) =$ uniform over the environment

- Find most likely position:

MAP=ML decoding

$$x_t = \underset{x}{\operatorname{argmax}} \left[P\left(\{s_{i,t}\} \middle| x \right) \times P_{prior}(x_t) \right]$$

Example: CA1 recording in freeling moving rats, 60 cm squared box, 34 cells
Place field: 400 squared bins per cell (discretized x − in 2D)
(data from K. Jezek)

Average error (Dt = 120 ms): 12 cm …

- Some regions in space are barely covered by place fields …
- Synthetic data study: error on positional decoding decreases as (nb. cells)$^{-1/2}$

Decoding of position from HPC activity

- Continuity prior over trajectory: $P_{prior}(\{x_t\}) \propto \prod_t \exp\left[-(x_t - x_{t-\Delta t})^2 / (2v^2\Delta t^2)\right]$

  Here, v = typical velocity (hyper-parameter, should be chosen with care)

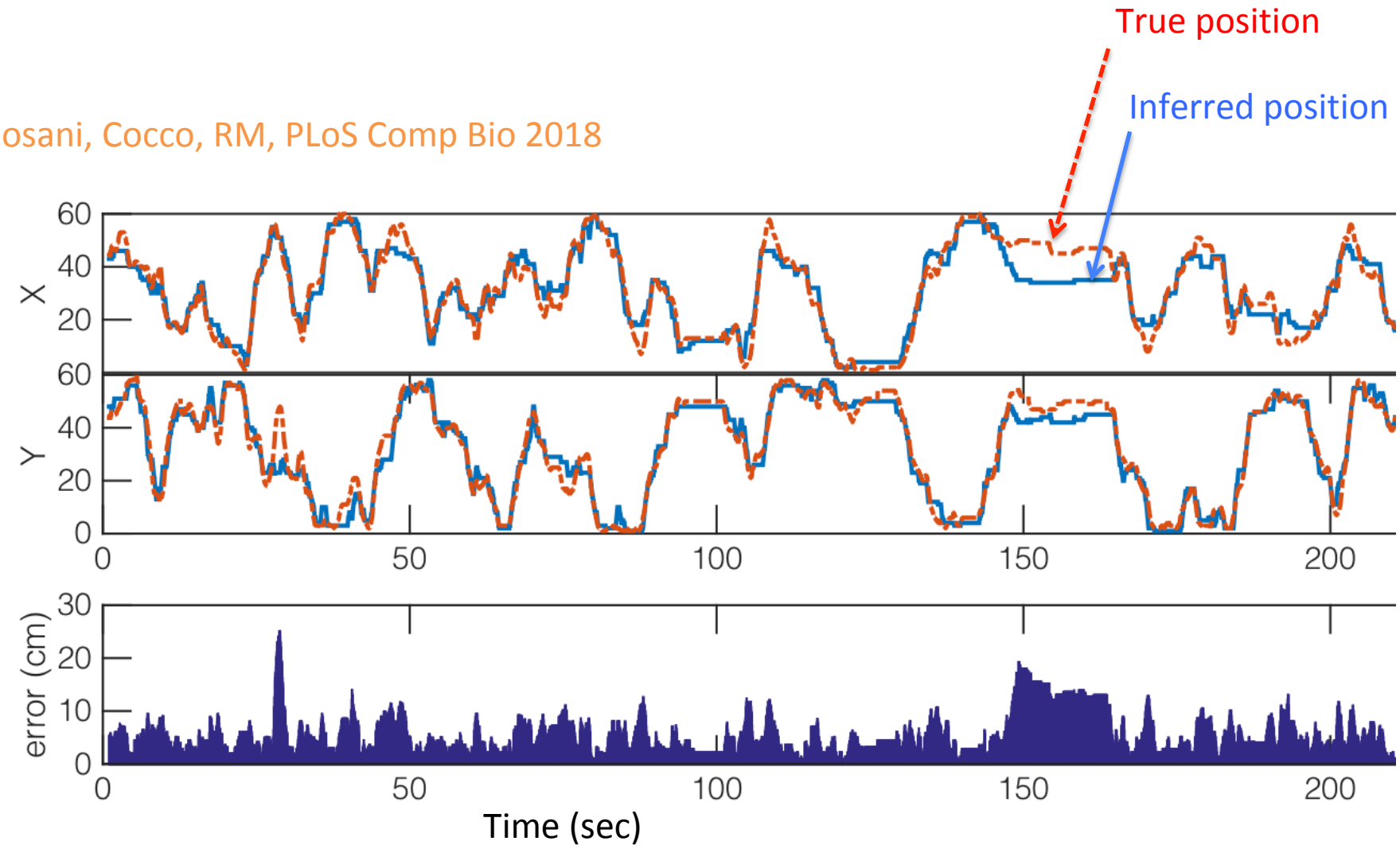
- Find most likely position:

MAP decoding $$\{x_t\}_{t=1,\ldots,T} = \text{argmax}\left[\prod_{t=1,\ldots,T} P\left(\{s_{i,t}\}\middle|x_t\right) \times P_{prior}\left(\{x_t\}\right)\right]$$

  - Brute force decoding impossible: exponential in *T*
  - Dynamic programming is efficient: linear in *T*, see Thursday

Decoding of position from HPC activity

Posani, Cocco, RM, PLoS Comp Bio 2018



Average error (Dt = 120 ms): 5.5 cm (compared to 12 cm)

Regression

Problem: Identify explanations/correlations between observations

Example: y_i^t = set of *N* behavioral/sensory correlates (e.g. position,..) at time t

x^t = activity of one given neuron at time t

Hypothesis: $x^t = \sum_{i=1,\ldots,N} a_i \; y_i^t + z^t$ with z^t = noise of mean 0, variance σ^2

We want to infer *N* parameters a_i from *T* observations. First, let us assume that *T* >> *N* (many data, good setting … : we do not need prior information, i.e. we assume uniform priors)

Likelihood : $$P\left(\left\{x^t\right\}\middle|\left\{a_i, y_i^t\right\}\right) = \prod_t \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x^t - \sum_i a_i \; y_i^t\right)^2\right)$$

Regression

Maximum Likelihood decoding = Least square method!

Minimize $$\sum_t \left(x^t - \sum_i a_i \; y_i^t \right)^2 = \sum_t \left(x^t \right)^2 - 2 \sum_i a_i b_i + \sum_{i,j} a_i C_{i,j} a_j$$

with $$b_i = \sum_t x^t y_i^t , C_{i,j} = \sum_t y_i^t y_j^t$$

Extremization condition : $$\frac{\partial}{\partial a_i}(\ldots) = 0 = -2 b_i + 2 \sum_j C_{i,j} a_j$$

Formal solution : $$\boxed{a_i = \sum_j \left(C^{-1} \right)_{i,j} b_j}$$

Statistical issues : (1) Inverse of matrix is not always well conditioned (not robust against measurement noise)

(2) Does not exist if $T < N$

Regularization

Idea : include prior distribution over coefficients can help!

L_2 prior : $$P\left(\{a_i\}\right) = \prod_i \frac{1}{\sqrt{2\pi\varepsilon^2}} \exp\left(-\frac{a_i^2}{2\varepsilon^2}\right)$$

Hyper-parameter

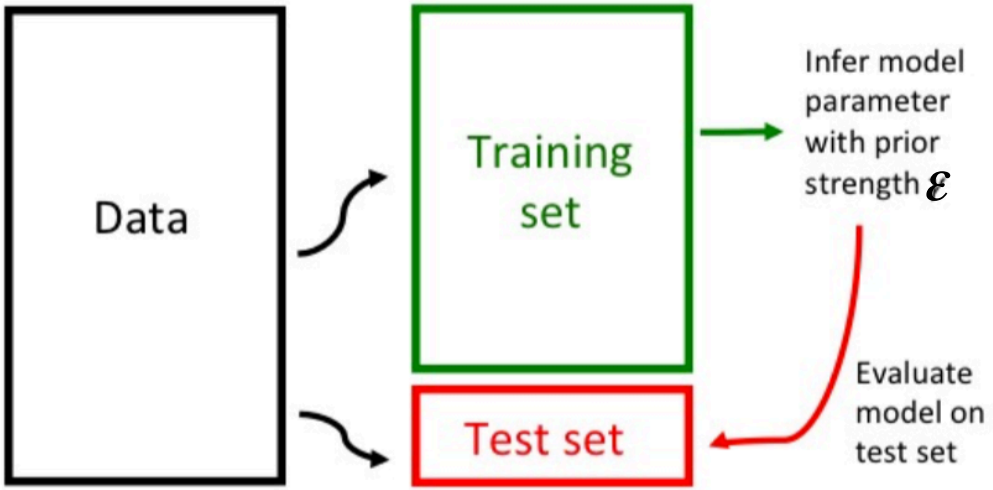
Maximization of posterior is equivalent to minimization of modified least square:

$$\frac{1}{2\sigma^2}\sum_t \left(x^t - \sum_i a_i \; y_i^t\right)^2 + \frac{1}{2\varepsilon^2}\sum_i a_i^2$$

Identical to previous complication with : $$C_{i,j} \rightarrow \tilde{C}_{i,j} = \sum_t y_i^t y_j^t + \frac{\sigma^2}{\varepsilon^2} Id_{i,j}$$

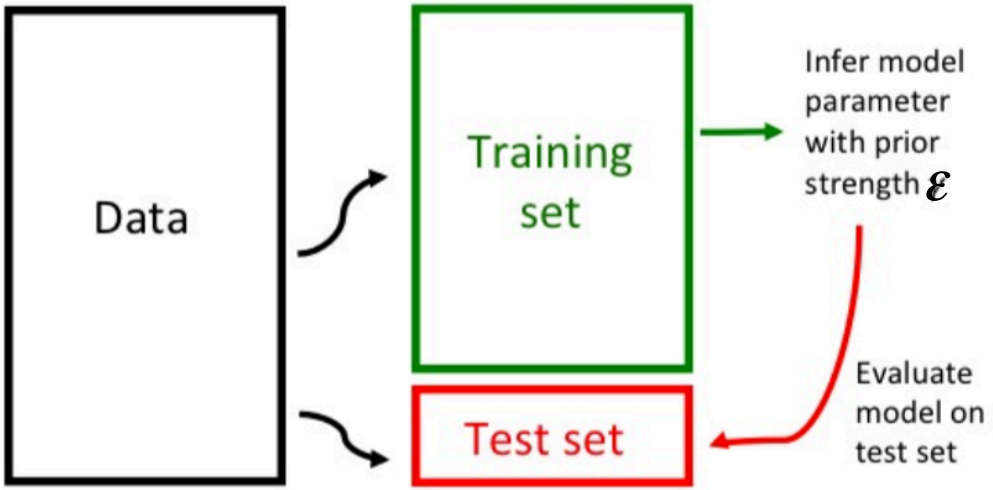
Well-defined and conditioned matrix, exists even for small *T* !

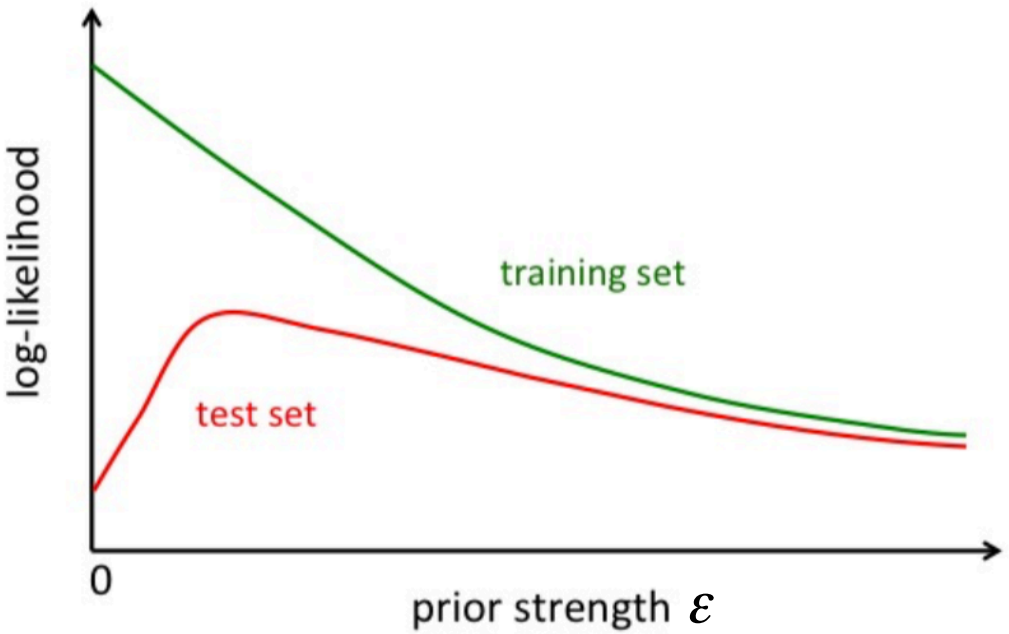
How to choose the value of hyper-parameters?



ε

How to choose the value of hyper-parameters?



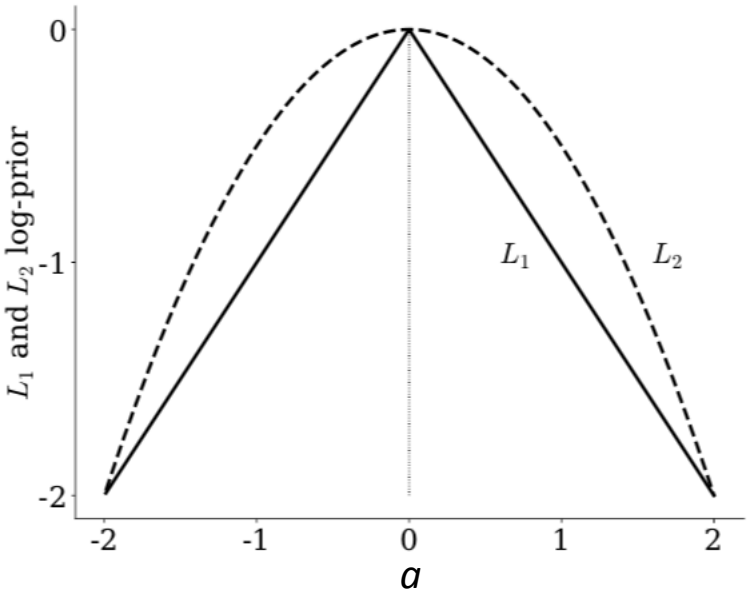


Regularization and sparse regressions

Idea : adequate prior distribution may favor the presence of many zero coefficients -> good for interpretability!

L_1 prior : $$P\left(\{a_i\}\right) = \prod_i \frac{1}{\varepsilon} \exp\left(-\frac{|a_i|}{\varepsilon}\right)$$

Hyper-parameter



Maximization of posterior is equivalent to minimization of

$$\frac{1}{2\sigma^2} \sum_t \left(x^t - \sum_i a_i \; y_i^t \right)^2 + \frac{1}{\varepsilon} \sum_i |a_i|$$

No simple analytical solution, but convex therefore easy to minimize!

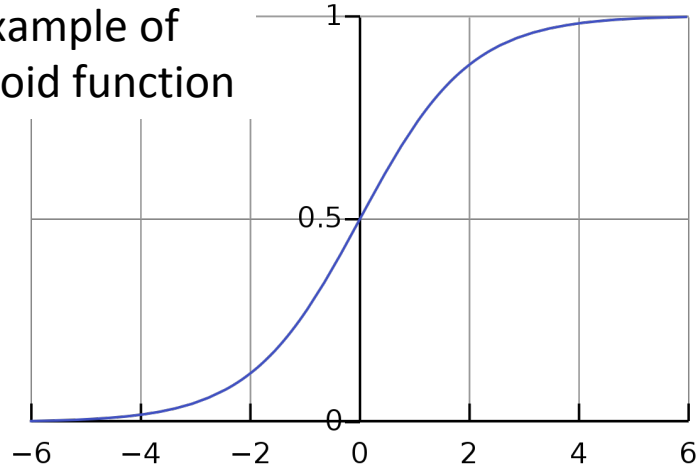
Case of class variables: logistic regression

Problem: sometimes x is not real-valued but corresponds to categories

Simplest binary setting : $x = 0,1$

$$x = \sum_{i=1,\ldots,N} a_i \; y_i + z \quad \rightarrow \quad P(x=1) = \Phi\left(\sum_{i=1,\ldots,N} a_i \; y_i \right), P(x=0) = 1 - P(1)$$

Example of sigmoid function

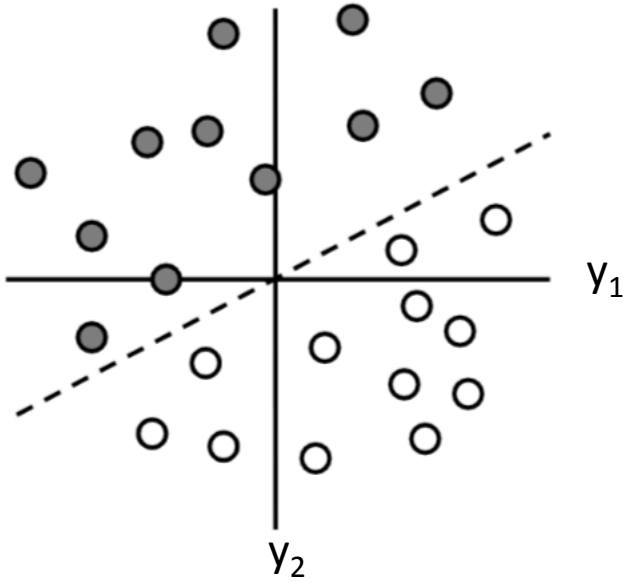


& offset: $\Phi\left(a_0 + \sum_{i=1,\ldots,N} a_i \; y_i \right)$

<u>Exercise 3:</u> write down Bayes formula for logistic regression

Classification

Logistic regression = linear classification (doubly linear, in coefficients and in data)

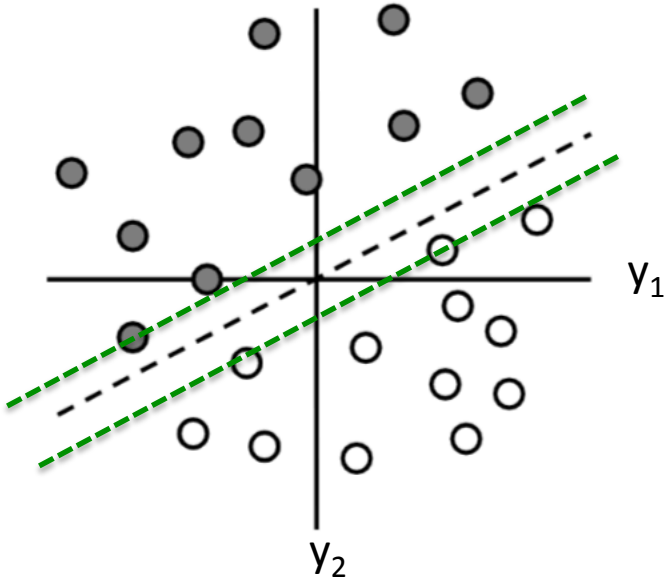


$$x = \theta\left(a_0 + \sum_i a_i y_i\right)$$

$$\theta(u) = 0 \ \textit{if} \ \ u \le 0, 1 \ \textit{if} \ \ u > 0$$

Classification

Logistic regression = linear classification (doubly linear, in coefficients and in data)



$$x = \theta\left(a_0 + \sum_i a_i y_i\right)$$

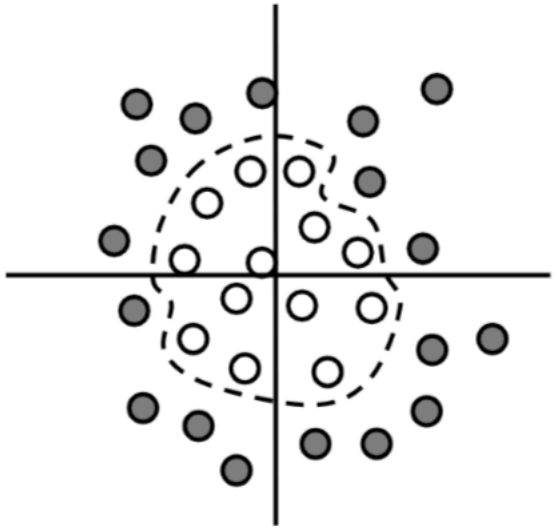
$$\theta(u) = 0 \ \textit{if} \ \ u \le 0, 1 \ \textit{if} \ \ u > 0$$

a.k.a. perceptron, support vector machine (maximal margin classifier)

Classification

Kernel methods = linear in coefficients, not linear in data

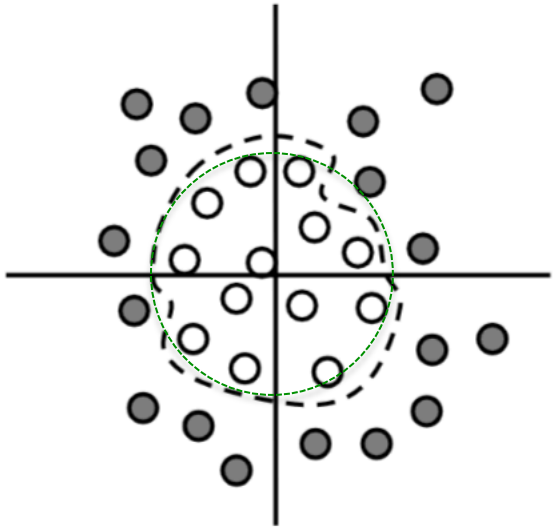
$$x = \theta\left(a_0 + \sum_i a_i y_i + \sum_{i \le j} a_{i,j} y_i y_j + \ldots \right)$$



Classification

Kernel methods = linear in coefficients, not linear in data

$$x = \theta\left(a_0 + \sum_i a_i y_i + \sum_{i \le j} a_{i,j} y_i y_j + \ldots\right)$$



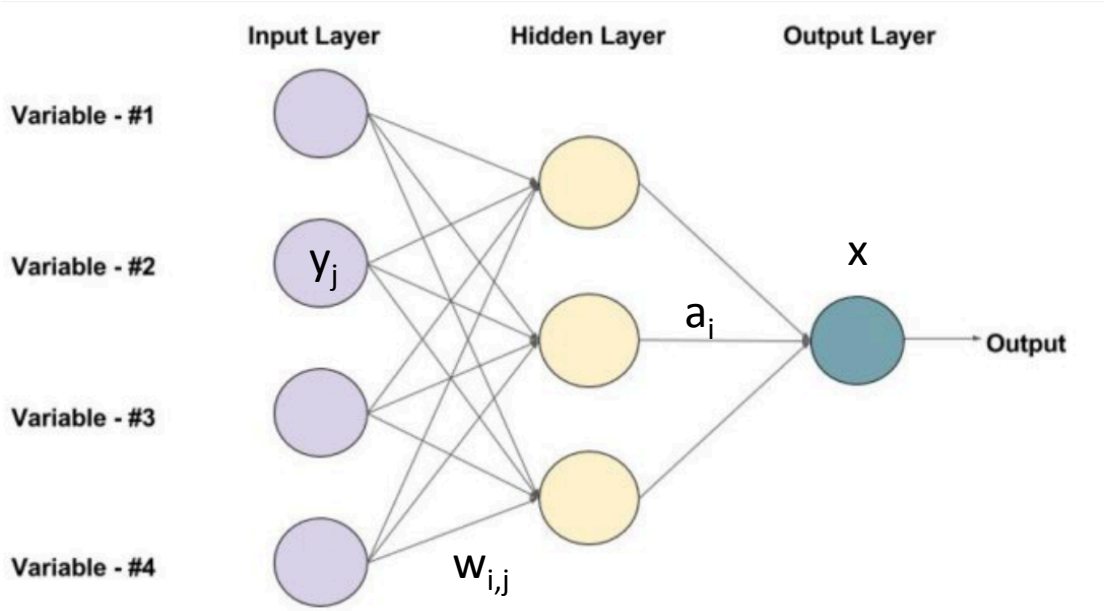
ex: $x = \theta\left(-R^2 + y_1 y_1 + y_2 y_2\right)$

$$\left(a_0 = -R^2, a_{1,1} = a_{2,2} = 1, a_i = 0, \ldots\right)$$

Fast : polynomial in nb. and in dimension of data

Classification

Multi-layer neural networks = Kernel methods, with kernel learned from data



$$x = \theta\left(\sum_i a_i \; g_i\left(\sum_j w_{i,j} y_j\right)\right)$$

Exercises

Exercise 1: *check that conditional probability is normalized!*

$$p_2\left(x_2 \middle| x_1\right) \equiv \frac{p(x_1, x_2)}{p_1(x_1)}$$

Exercises

Exercise 1: *check that conditional probability is normalized!*

$$p_2\left(x_2\middle|x_1\right) \equiv \frac{p(x_1,x_2)}{p_1(x_1)}$$

Answer:

$$\sum_{x_2} p_2\left(x_2\middle|x_1\right) = \sum_{x_2} \frac{p(x_1,x_2)}{p_1(x_1)} = \frac{\sum_{x_2} p(x_1,x_2)}{p_1(x_1)} = \frac{p_1(x_1)}{p_1(x_1)} = 1$$

Exercises

Exercise 2: *compute MI(box,cookie type) in bits*

Box *A*

20 plain cookies
+
20 chocolate cookies

Box *B*

10 plain cookies
+
30 chocolate cookies

Exercises

Exercise 2: *compute MI(box,cookie type) in bits*

$$MI(box, cookie\ type) = \sum_{x_1, x_2} p(x_1, x_2)\ \log\left[\frac{p(x_1, x_2)}{p_1(x_1)\, p_2(x_2)}\right]$$

$p(x_1, x_2)$	A	B
$Plain$	$\frac{1}{4}$	$\frac{1}{8}$
$Choco$	$\frac{1}{4}$	$\frac{3}{8}$

→ Gives marginal distribution by summing columns or rows

$$MI = \frac{1}{4}\log_2\left[\frac{\frac{1}{4}}{\frac{1}{2}\times\frac{3}{8}}\right] + \frac{1}{4}\log_2\left[\frac{\frac{1}{4}}{\frac{1}{2}\times\frac{5}{8}}\right] + \frac{1}{8}\log_2\left[\frac{\frac{1}{8}}{\frac{1}{2}\times\frac{3}{8}}\right] + \frac{3}{8}\log_2\left[\frac{\frac{3}{8}}{\frac{1}{2}\times\frac{5}{8}}\right] \approx 0.05\ bit$$