Probabilistic models for neural data: From single neurons to population dynamics

NEUROBIO 316QC

Jan Drugowitsch

jan_drugowitsch@hms.harvard.edu

Session 3: Generalized linear models #1

Today

Q&A about previous session

Paper discussion (~1h)

Introducing generalized linear models (remaining time)

Overview

Bernoulli distribution, the exponential family & conjugate priors

- Bernoulli distribution as spike generation model
- Conjugate priors for the Bernoulli distribution
- The exponential family of probability distributions

Linear classification models

- Generative vs. discriminative models
- Logistic regression: sigmoidal Bernoulli probability
- Inhomogeneous Poisson process: sequence of Bernoulli events
- Generalized linear models and canonical activation function

Overview

Bernoulli distribution, the exponential family & conjugate priors

- Bernoulli distribution as spike generation model
- Conjugate priors for the Bernoulli distribution
- The exponential family of probability distributions

Linear classification models

- Generative vs. discriminative models
- Logistic regression: sigmoidal Bernoulli probability
- Inhomogeneous Poisson process: sequence of Bernoulli events
- Generalized linear models and canonical activation function

The Bernoulli distribution

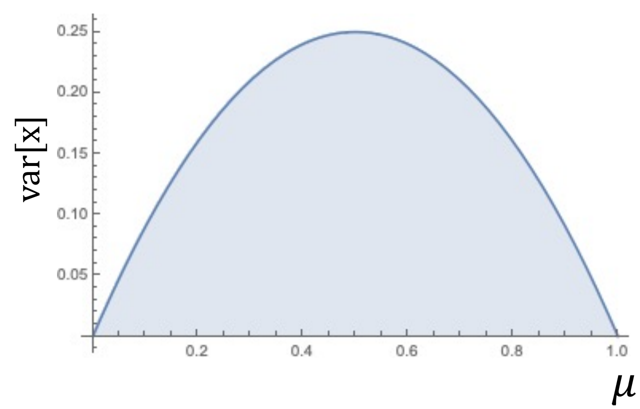
To spike ($x = 1$) or not to spike ($x = 0$)?

$$\left.\begin{array}{l} p(x = 1|\mu) = \mu \\ p(x = 0|\mu) = 1 - \mu \end{array}\right\} p(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

Moments

$$\mathrm{E}[x] = \mu$$

$$\mathrm{var}[x] = \mu(1 - \mu)$$



certain ⟷ uncertain ⟷ certain

Conjugate priors

Posterior has same functional form (e.g., Gaussian) as prior

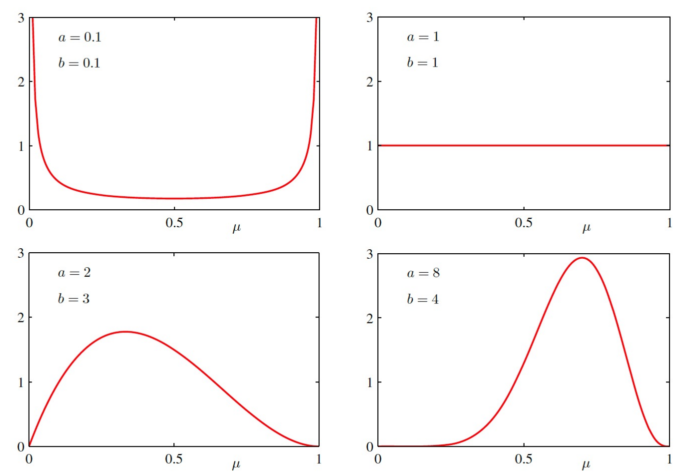
same functional form

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

depends on likelihood

Example: Bernoulli distribution

$$\begin{aligned}
\log p(\mu|x) &= \log p(x|\mu) + \log p(\mu) + \text{const.} \\
&= x \log \mu + (1-x)\log(1-\mu) + \log p(\mu) + \text{const.} \\
&= x \log \mu + (1-x)\log(1-\mu) + \underbrace{(\alpha-1)\log \mu + (\beta-1)\log(1-\mu)}_{\text{Beta distribution prior } \mathrm{B}(\mu|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}} + \text{const.} \\
&= \underbrace{(x+\alpha-1)\log \mu + (1-x+\beta-1)\log(1-\mu)}_{\text{Beta distribution posterior } \mathrm{B}(\mu|\alpha+x,\beta+1-x)} + \text{const.}
\end{aligned}$$



For multiple observations, $x_1, x_2, \ldots$:

$$\tilde{\alpha} = \alpha + \sum_n x_n \qquad \tilde{\beta} = \beta + \sum_n (1 - x_n)$$

The exponential family of distributions

base measure natural parameters

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x}))$$

normalizer sufficient statistics

Base measure: density $p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})$ for $\boldsymbol{\eta} = \mathbf{0}$

Normalizer: ensures $\int p(\boldsymbol{x}|\boldsymbol{\eta})\mathrm{d}\boldsymbol{x} = 1$

Sufficient statistics: consider i.i.d. data $x_1, \ldots, x_N$

$$p(\boldsymbol{x}_{1:N}|\boldsymbol{\eta}) \propto_\eta g(\boldsymbol{\eta})^N \exp\left(\boldsymbol{\eta}^T \underbrace{\sum_{n=1}^{N} u(\boldsymbol{x}_n)}_{\text{summarizes what } \boldsymbol{x}_{1:N} \text{ tells us about } \boldsymbol{\eta}}\right)$$

Fisher-Darmois-Koopman-Pitman theorem:
(finite) sufficient statistics only exist for exponential family

Conjugate prior: follows directly from form of exponential familty

Instances: Bernoulli, Beta, Gaussian, Poisson, Laplace, …

Mixture models are not members of the exponential family

Interim summary: Bernoulli distribution & exponential family

Bernoulli distribution: models binary events (e.g., spike/no spike in small time window)

Conjugate priors: posteriors have same functional form

Including information from likelihood: updating distribution parameters

Beta distribution is conjugate prior for Bernoulli probability

Exponential family: large family of probability distributions

Only family that can efficiently summarize information from data (i.e., finite suff. stats)

Members: Bernoulli Gaussian, Poisson, Exponential, Laplace, …

Overview

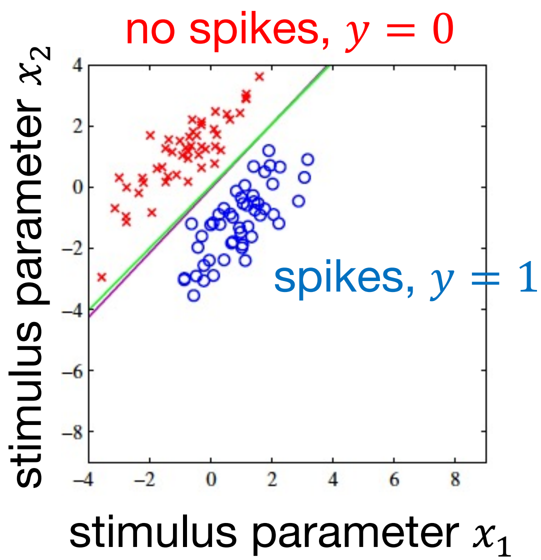
Bernoulli distribution, the exponential family & conjugate priors

- Bernoulli distribution as spike generation model
- Conjugate priors for the Bernoulli distribution
- The exponential family of probability distributions

Linear classification models

- Generative vs. discriminative models
- Logistic regression: sigmoidal Bernoulli probability
- Inhomogeneous Poisson process: sequence of Bernoulli events
- Generalized linear models and canonical activation function

Linear models for classification



PRML, Fig. 4.4

Standard linear model?

$$y = \boldsymbol{w}^T \boldsymbol{x} + \eta$$

Better: non-linear linear model, s.t. $y \in \{0,1\}$

$y = 1$ if $\boldsymbol{w}^T \boldsymbol{x} + \eta > \theta$, $y = 0$ otherwise

leads to

$$p(y = 1|\boldsymbol{x}, \boldsymbol{w}) = f(\boldsymbol{w}^T \boldsymbol{x})$$

Generative model: specify $p(\boldsymbol{x}|y, \boldsymbol{w})$, i.e., distribution over stimuli to generate spike

$$p(y|\boldsymbol{x}, \boldsymbol{w}) \propto p(\boldsymbol{x}|y, \boldsymbol{w})p(y)$$

provides full model of stimulus distribution (many parameters, might be hard to learn)

Discriminative model: specify $p(y|\boldsymbol{x}, \boldsymbol{w})$, i.e., spike probabilities given stimulus
Does not a-priori model stimulus distribution (fewer parameters, might be easier to learn)

They are related by Bayes' rule! (one implicitly determines part of the other)

Discriminative linear models

Monotonically increasing activation/link function $f(\cdot)$: $y = 1$ if $\underbrace{\boldsymbol{w}^T\boldsymbol{x} + \eta} > \tilde{\theta}$

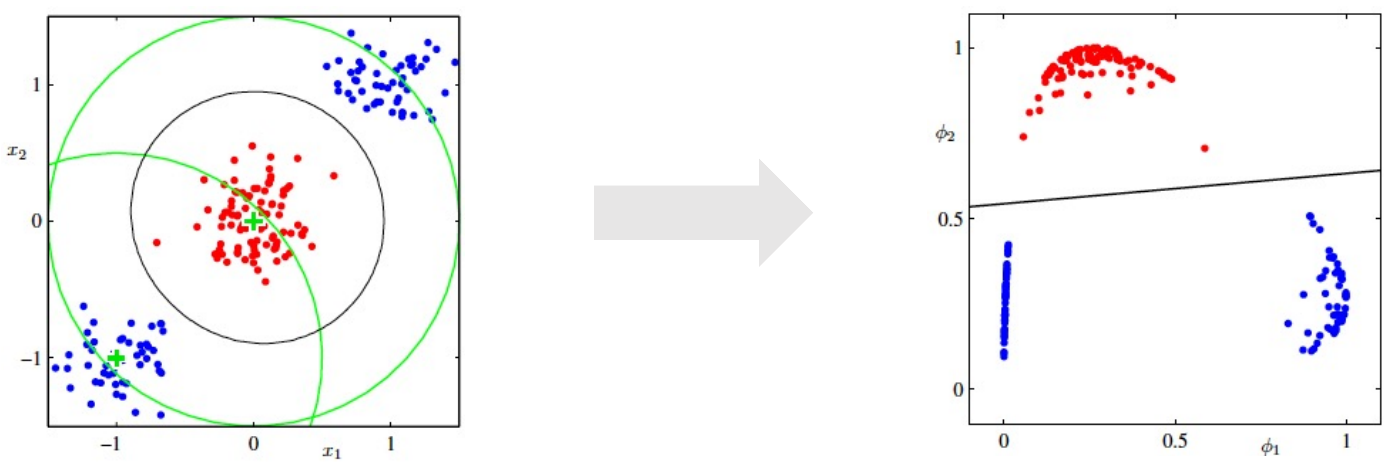
(noisy) **linear threshold in $\boldsymbol{x}$**

(Non-linear) **linear in parameters**, not ‘inputs’

as for linear models, can use non-linear functions $\boldsymbol{\Phi}(\boldsymbol{x})$ of input

$$p(y = 1|\boldsymbol{x}, \boldsymbol{w}) = f(\boldsymbol{w}^T\boldsymbol{\Phi}(\boldsymbol{x}))$$

Non-linear functions can make (linear) non-discriminable data discriminable



Logistic regression

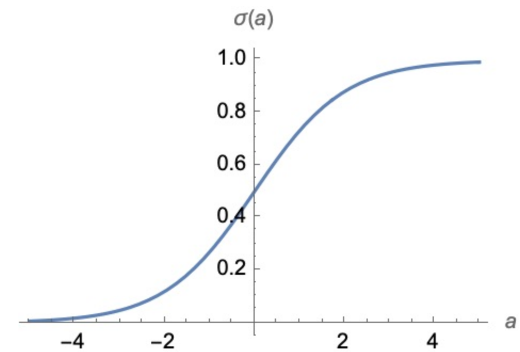
Assumes log-odds linear in $\boldsymbol{x}$ (or $\boldsymbol{\Phi}(\boldsymbol{x})$)

log-posterior ratio   log-likelihood ratio (LLR)

$$\log\frac{p(y=1|\boldsymbol{x},\boldsymbol{w})}{p(y=0|\boldsymbol{x},\boldsymbol{w})} = \log\frac{p(\boldsymbol{x}|y=1,\boldsymbol{w})}{p(\boldsymbol{x}|y=0,\boldsymbol{w})} + \log\frac{p(y=1)}{p(y=0)} \equiv a(\boldsymbol{x},\boldsymbol{w})$$

leads to

$$p(y=1|\boldsymbol{x},\boldsymbol{w}) = \underbrace{\frac{1}{1+e^{-a(\boldsymbol{x},\boldsymbol{w})}}}_{\text{logistic sigmoid}} \equiv \sigma(a(\boldsymbol{x},\boldsymbol{w}))$$



Assuming **Gaussian likelihoods**, $p(\boldsymbol{x}|y) = N(\boldsymbol{x}|\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ $a(\boldsymbol{x},\boldsymbol{w}) = \underbrace{-\tfrac{1}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}}_{\boldsymbol{w}^T}\boldsymbol{x} + \text{const.}$

$$p(y=1|\boldsymbol{x},\boldsymbol{w}) = \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}}$$

In general: compatible with any generative model for which LLR $\sim \boldsymbol{w}^T\boldsymbol{x} + \text{const.}$

Moving to spike trains

Logistic regression: for individual spikes in small time bins (spike or no spike?)

Assume spike train $y_{1:N}$ with i.i.d. Bernoulli probabilities, $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t$

(very)small time bin

$$\log p(y_{1:N}|\boldsymbol{x}_{1:N}) = \sum_n y_n \log \lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t + \sum_n (1 - y_n) \log(1 - \lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t)$$

$\log(1 + a\delta t) \approx a\delta t$

$$\approx \sum_n y_n \log \lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t - \sum_n (1 - y_n)\lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t$$

$$= \sum_n y_n (\log \lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t + \lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t) - \sum_n \lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t$$

$|\log \lambda\delta t| \gg |\lambda\delta t|$

$\log y_n! = 0$

$$\approx \sum_n y_n \log \lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t - \sum_n \lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t - \sum_n \log y_n!$$

$$= \underbrace{\log \prod_n \mathrm{Pois}(y_n|\, \lambda(\boldsymbol{x}_n, \boldsymbol{w})\delta t)}_{\text{time-discretized inhomogeneous Poisson process}}$$

Properties of the (in)homogeneous Poisson process

Derivable from i.i.d. Bernoulli probabilities:
spiking only depends on instantaneous rate λ, otherwise independent across time

More generally: spike counts of any two non-overlapping time periods are independent

Spike count within any time interval is distributed by a Poisson distribution

Within any small time-interval δt with instantaneous rate λ_t:
Spike count N satisfies $\mathrm{E}[N] = \lambda_t \delta t$ and $\mathrm{var}[N] = \lambda_t \delta t$; Fano factor $= \frac{\mathrm{var}[N]}{\mathrm{E}[N]} = 1$

Special case: homogeneous Poisson process ($\boldsymbol{\lambda_t = \mathrm{const.}}$)

Within any time interval Δ: spike count N has distribution $\mathrm{N} \sim \mathrm{Pois}(\lambda\Delta)$
Expected number of spikes in that interval are $\mathrm{E}[N] = \lambda\Delta$

Inter-spike interval have exponential distribution

The exponential family and canonical activation functions

Linear regression: Gaussian noise + identity activation function, $f(x) = x$

$$\log p(y_{1:N} | \boldsymbol{x}_{1:N}, \boldsymbol{w}) \propto -\sum_n (\boldsymbol{w}^T \boldsymbol{\Phi}(\boldsymbol{x}_n) - y_n)^2$$

gradient for single n:

$$\nabla_w \log p(y_n | \boldsymbol{x}_n, \boldsymbol{w}) \propto -(\boldsymbol{w}^T \boldsymbol{\Phi}(\boldsymbol{x}_n) - y_n)\boldsymbol{\Phi}(\boldsymbol{x}_n)$$

estimate target

Logistic regression: Bernoulli noise + sigmoidal activation function, $f(x) = \sigma(x)$

$$\log p(y_{1:N} | \boldsymbol{x}_{1:N}, \boldsymbol{w}) = \sum_n y_n \log \sigma\left(\boldsymbol{w}^T \boldsymbol{\Phi}(\boldsymbol{x}_n)\right) + (1 - y_n) \log\left(1 - \sigma\left(\boldsymbol{w}^T \boldsymbol{\Phi}(\boldsymbol{x}_n)\right)\right)$$

gradient for single n:

$$\nabla_w \log p(y_n | x_n, w) = -\left(\sigma\left(\boldsymbol{w}^T \boldsymbol{\Phi}(\boldsymbol{x}_n)\right) - y_n\right) \boldsymbol{\Phi}(\boldsymbol{x}_n)$$

estimate target

Poisson regression: Poisson noise + exponential activation function, $f(x) = e^x$

Same gradient form: supports the same Iterated Recursive Least Squares (IRLS) algorithm

Generally: likelihood $p(y_n | x_n, w)$ determines *canonical* activation function (RPML 4.3.6)

Interim summary

Generative and discriminative linear models are related by Bayes' rule

Discriminative logistic regression assumes generative log-likelihood ratio (LLR) linear in $\boldsymbol{w}$

Inhomogeneous Poisson processes = Bernoulli spikes with probability rate $\times\ \delta t$

Assume spikes independent across time (conditional on rate)

Exponential family distributions have canonical activation function that support IRLS

Poisson regression has exponential activation function

Summary

Bernoulli distribution: simple model for instantaneous spiking

Assuming i.i.d. Bernoulli spikes in small time windows δt:
Inhomogeneous Poisson process, instantaneous rate = spike probability / δt

Exponential family: large family of probability distributions with convenient properties
- conjugate prior: easily parameter updates
- ability to summarize data in finite sufficient statistics for parameter inference

Linear models for classification
Generative vs. discriminative models
Logistic regression: discriminative model assuming LLR linear in $\boldsymbol{\Phi}(\boldsymbol{x})$

Generalized linear models
Likelihood function & canonical activation function
Includes linear regression, logistic regression, Poisson regression, etc.

Until next week

Read paper and prepare presentation (see notes for Session 4)

Read statistical methods sections (see notes for Session 4)

Next session

Q&A for previous session (~15min)

Paper discussions (~1h)

Generalized linear models, graphical models & state space models (~30min)