Lecture 2

Supervised Learning
Part 1

TAs
Ajinkya Deogade
Amrita Singh
Diptodip Deb
Nils Eckstein
Virginia Rutten

Srini Turaga
turagas@janelia.hhmi.org

1. Key elements of SL/ML
   - data "x", "y"
   - prediction function $\hat{y} = f(x, \theta)$
   - loss function $\mathcal{L}(y, \hat{y})$
   - optimization / training algorithm
   - regularizer (optional) $R(\theta)$
2. Linear regression
3. Classification as generalized linear modeling
4. Loss functions can be derived from probilistic interpretations

The what and why of supervised learning

parameters

$$\hat{y} \xleftarrow{f(x,\theta)} x$$

$\hat{y}$	x
number	vector
category	image
vector	sequence
image	all of the above
sequence	

Black-box vs Interpretable

The what and how of supervised learning

1. x_1, y_1
x_2, y_2
x_3, y_3
$\vdots$
x_N, y_N

Examples/data

2. $\hat{y}_i = f(x_i, \theta)$

prediction function

3. $\mathcal{L}(y_i, \hat{y}_i)$

scoring/objective/loss function

4. $R(\theta)$

regularization function

(optional)

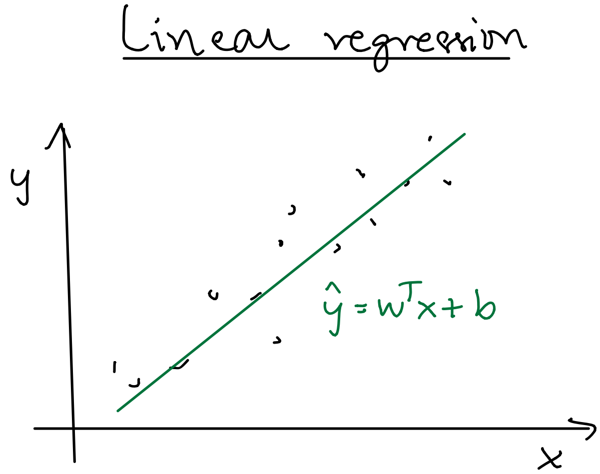
5. $$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(y_i, \hat{y}_i) + R(\theta)$$
$$= \min_{\theta} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i, \theta)) + R(\theta)$$

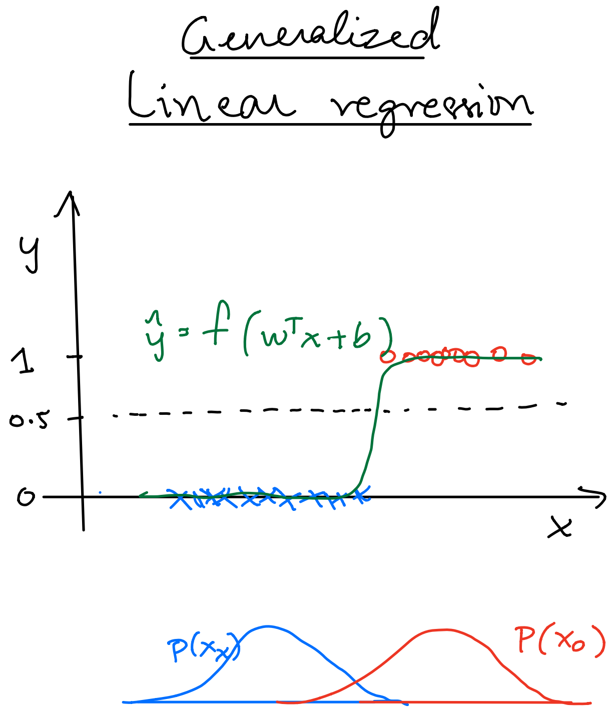
optimization algorithm

Generalized Linear Models

$u = w^T x + b$ ⟵ linear $\theta = \{w, b\}$

$\hat{y} = f(u)$ ⟵ nonlinear





Technical note:

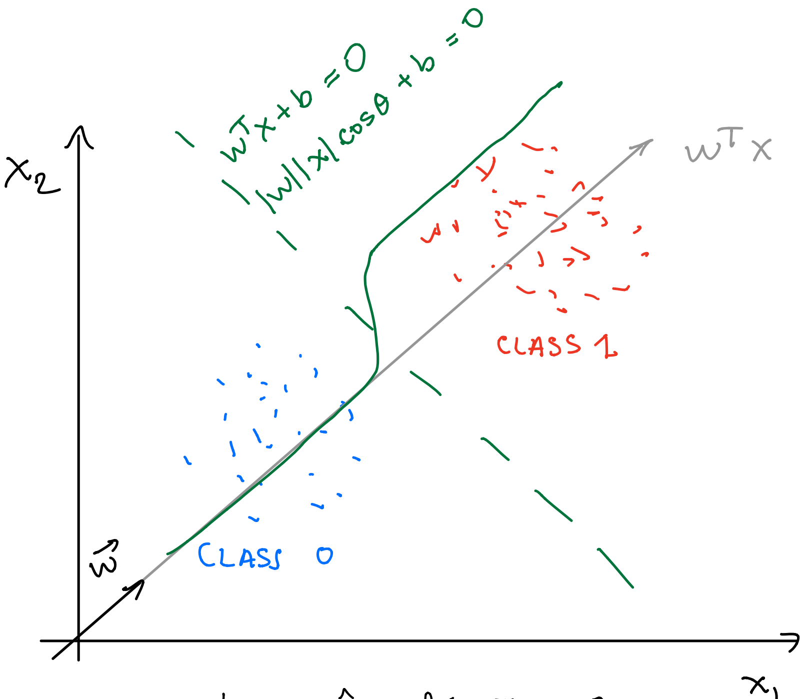
$$\hat{y} = f(w^T x + b) \quad \longleftarrow f(u) : \text{nonlinearity}$$

Equivalently

$$f^{-1}(\hat{y}) = w^T x + b \quad \longleftarrow f^{-1}(y) : \text{link function}$$

Logistic regression

binary classification

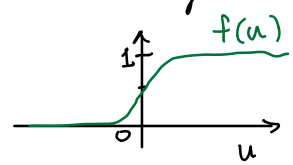


$u = w^T x + b$

$\hat{y} = f(u)$

$$f(u) = \frac{1}{1 + e^{-u}}$$

logistic sigmoid nonlinearity

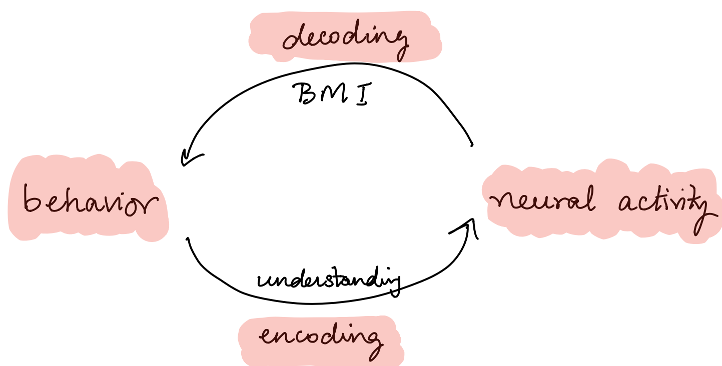


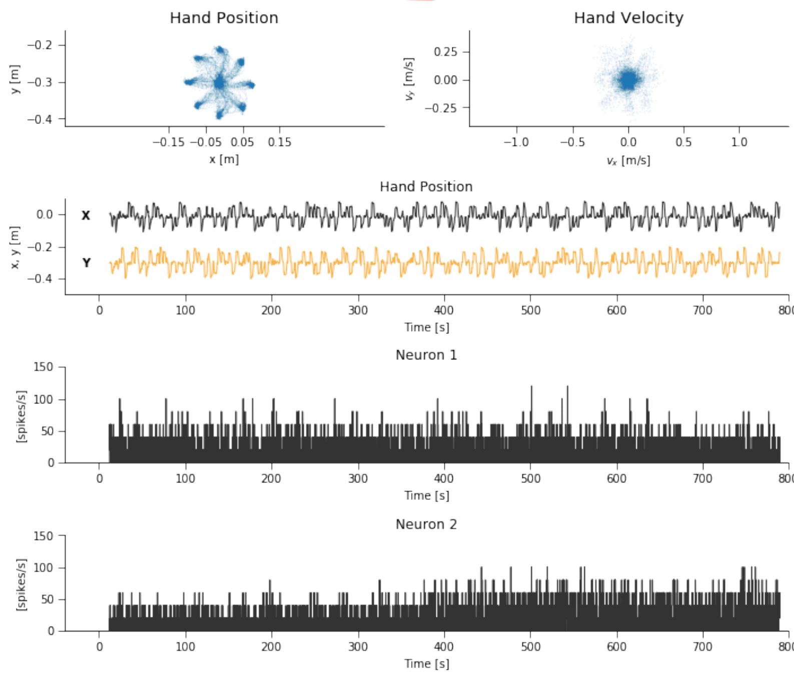
$p(y=1 \mid x) = \hat{y} = f(w^T x + b)$

Loss function : $-\sum_i y_i \log \hat{y}_i + (1-y_i) \log (1-\hat{y}_i)$

$= - \log p(y_i \mid \hat{y}_i = f(w^T x + b))$

Neural data analysis





Poisson regression

firing rate of a neuron at time t

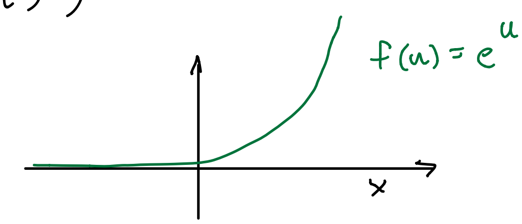
$$\hat{y}(t) = f(\underbrace{w^T}_{\text{filter}} \underbrace{x(t)}_{\text{stimulus}} + b)$$

non-negative

actual spikes are non-negative integers

$$y(t) \sim \text{Poisson}(\hat{y}(t))$$

So choose $f(u) = e^u$



Loss function : $\sum_i \hat{y}_i - y_i \log \hat{y}_i$

$$= -\log P(y_i \mid \hat{y}_i = f(w^T x + b))$$

Sparse regression

how to find informative features

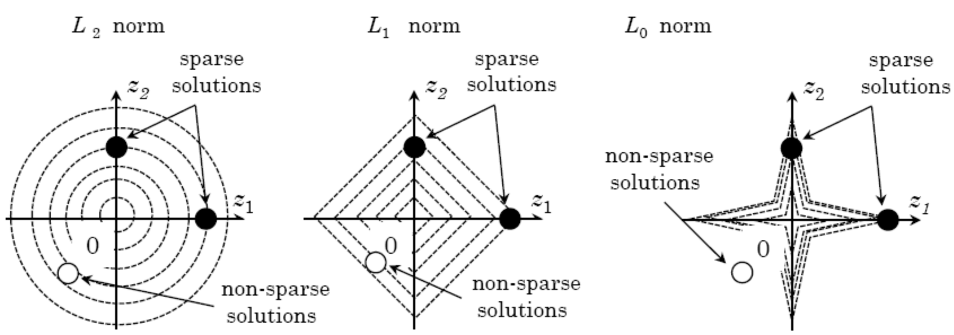
$$\hat{y} = w^T x + b$$
$$= w_1 x_1 + w_2 x_2 + \cdots + b$$

Optimize : $\sum_i \mathcal{L}(y_i, \hat{y}_i) + \lambda . R(w)$ penalty for using features

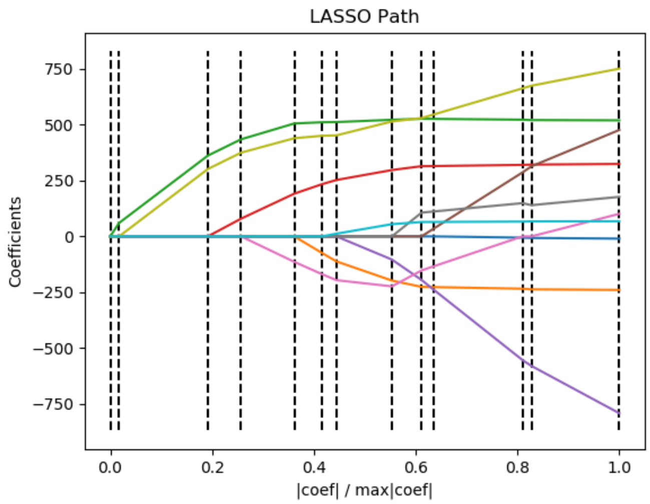
$$= \frac{1}{2} \sum_i (y_i - \hat{y}_i)^2 + \lambda . R(w)$$

Trade off between $R(w)$ & $\mathcal{L}(y, \hat{y})$

Standard choice : $R(w) = \sum_j |w_j|^1 = \|w\|_1$



Lasso : sparse L1 regularized regression



← increasing λ

→ decreasing sparsity

The what and why of supervised learning

$$\hat{y} \xleftarrow{f(x,\theta)} x$$

parameters

$\hat{y}$	x
number	vector
category	image
vector	sequence
image	all of the above
sequence	

Black-box vs Interpretable

The what and how of supervised learning

1. x_1, y_1
x_2, y_2
x_3, y_3
$\vdots$
x_N, y_N

Examples/data

2. $\hat{y}_i = f(x_i, \theta)$

prediction function

3. $\mathcal{L}(y_i, \hat{y}_i)$

scoring/objective/loss function

4. $R(\theta)$

regularization function

(optional)

5. $$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(y_i, \hat{y}_i) + R(\theta)$$
$$= \min_{\theta} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i, \theta)) + R(\theta)$$

optimization algorithm