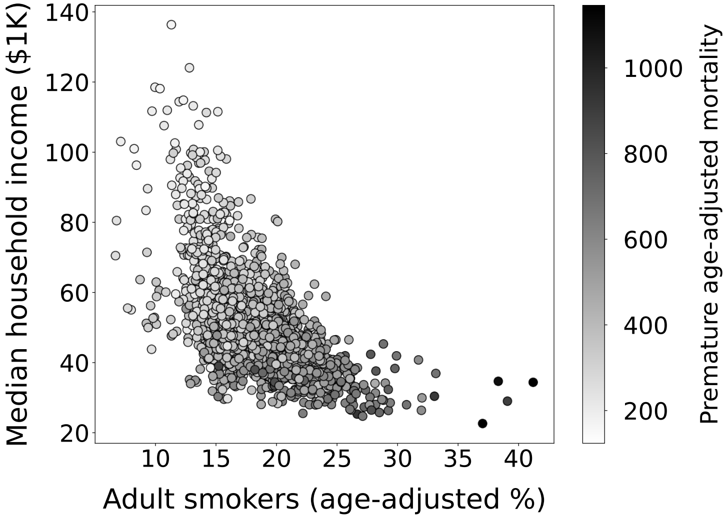
Logistic Regression

Probability and Statistics for Data Science

Carlos Fernandez-Granda

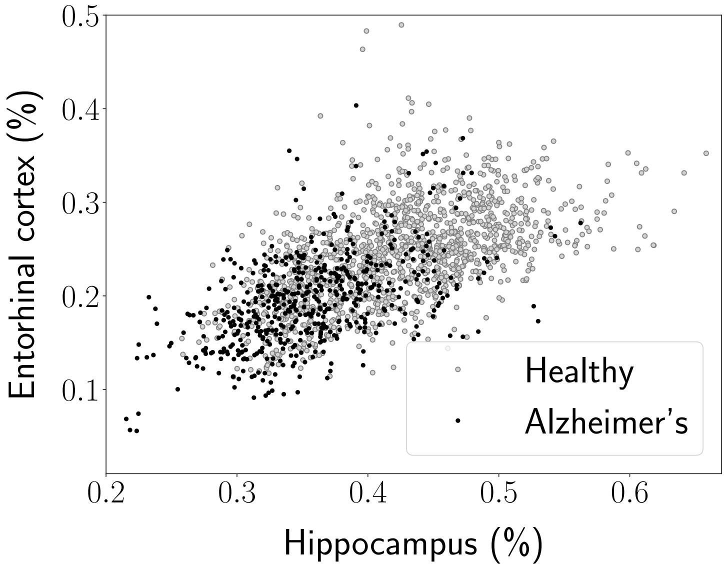
These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Regression



Median household income ($1K)
140
120
100
80
60
40
20
10
15
20
25
30
35
40
Adult smokers (age-adjusted %)
Premature age-adjusted mortality
1000
800
600
400
200

Classification



Classification

Data: (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n)

Each feature x_i is a d-dimensional vector (e.g. MRI scan)

The label y_i indicates the class (e.g. *Alzheimer's* or *healthy*)

Goal: Assign class to new data

Probabilistic modeling

Model features as random vector $\tilde{x}$ and label as random variable $\tilde{y}$

For new data vector x:

$$\hat{y} := \arg \max_{y \in \{1,2,\ldots,c\}} p_{\tilde{y} \,|\, \tilde{x}}(y \,|\, x)$$

Is classification easy?

Curse of dimensionality

Unless number of features (entries in $\tilde{x}$) is very small, it is impossible to estimate $p_{\tilde{y} \mid \tilde{x}}(y \mid x)$!

For m binary features we need to estimate 2^m conditional pmfs!

We need assumptions!

Generative vs discriminative approaches

Naive Bayes and Gaussian discriminant analysis are generative approaches

First estimate $p_{\tilde{x} \mid \tilde{y}}$ / $f_{\tilde{x} \mid \tilde{y}}$ and $p_{\tilde{y}}$, then apply Bayes' rule

Discriminative approaches estimate $p_{\tilde{y} \mid \tilde{x}}$ directly

Logistic regression

Discriminative *binary* classification method (two classes 0 and 1)

Goal: Use linear model to approximate $p_{\tilde{y} \mid \tilde{x}}(1 \mid x)$

First idea:

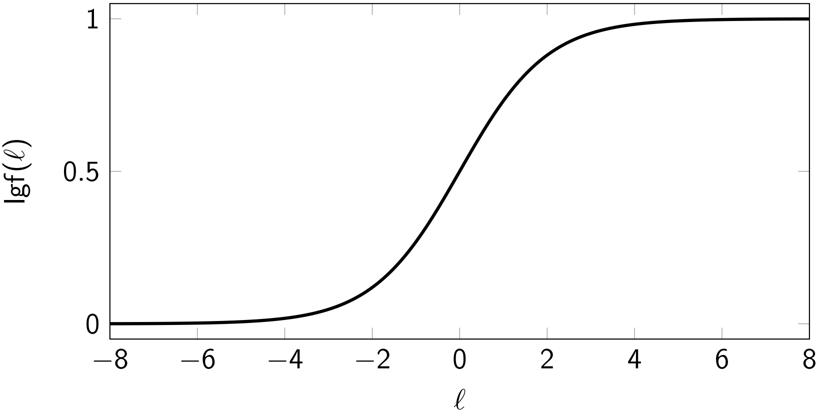
$$p_{\tilde{y} \mid \tilde{x}}(1 \mid x) = \beta^T x + \alpha$$

Problem: For most values of x *not a valid probability*

Solution:
Map linear function of features to [0,1] using link function

Logistic function

$$\mathsf{lgf}(\ell) := \frac{\exp(\ell)}{1+\exp(\ell)} = \frac{1}{1+\exp(-\ell)}$$

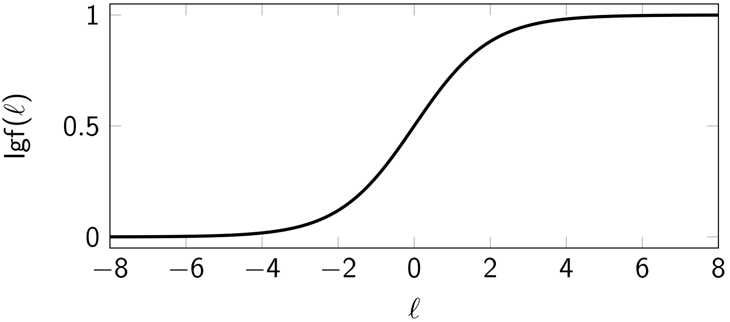


Logistic regression

Generalized linear model

$$\mathrm{P}\left(\tilde{y} = 1 \,|\, \tilde{x} = x\right) = \mathrm{lgf}(\beta^T x + \alpha)$$

Properties of the logistic function



1. Monotone
   $\beta^T x_1 + \alpha > \beta^T x_2 + \alpha \implies \mathrm{lgf}(\beta^T x_1 + \alpha) > \mathrm{lgf}(\beta^T x_2 + \alpha)$

2. Continuous
   $\beta^T x_1 + \alpha \approx \beta^T x_2 + \alpha \implies \mathrm{lgf}(\beta^T x_1 + \alpha) \approx \mathrm{lgf}(\beta^T x_2 + \alpha)$

3. Saturates: Large negative / positive inputs mapped to 0 / 1

Logits

The logit function is the inverse of the logistic function

Inputs to the logistic function are called logits

Odds

Ratio between probability of an event and probability of complement

If $\mathrm{P}(A) = 0.5$, odds are 1. If $\mathrm{P}(A) = 0.75$, odds are 3

For $\mathrm{P}(A) := p = \operatorname{lgf}(\ell)$

$$\text{odds} = \frac{p}{1-p} = \frac{\operatorname{lgf}(\ell)}{1-\operatorname{lgf}(\ell)} = \frac{\frac{\exp(\ell)}{1+\exp(\ell)}}{1-\frac{\exp(\ell)}{1+\exp(\ell)}} = \exp(\ell)$$

In logistic regression, log odds are affine function of the features

$$\log\left(\frac{\mathrm{P}\left(\tilde{y}=1 \,|\, \tilde{x}=x\right)}{1-\mathrm{P}\left(\tilde{y}=1 \,|\, \tilde{x}=x\right)}\right) = \beta^T x + \alpha$$

Parameter estimation

How do we estimate the logistic-regression parameters β and α?

Data: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)

Model the ith feature and label as the random variables $\tilde{x}_i$ and $\tilde{y}_i$

Maximize the conditional likelihood of the labels given the features

Likelihood

Assumption 1:

Labels are conditionally independent given the features

Assumption 2:

$\tilde{y}_i$ is conditionally independent from $\{\tilde{x}_m\}_{m \neq i}$ given $\tilde{x}_i$

$$
\begin{aligned}
\mathcal{L}_{XY}(\alpha, \beta) &:= \mathrm{P}\left(\tilde{y}_1 = y_1, \ldots, \tilde{y}_n = y_n \,|\, \tilde{x}_1 = x_1, \ldots, \tilde{x}_n = x_n\right) \\
&= \prod_{i=1}^{n} \mathrm{P}\left(\tilde{y}_i = y_i \,|\, \tilde{x}_1 = x_1, \ldots, \tilde{x}_n = x_n\right) \\
&= \prod_{i=1}^{n} \mathrm{P}\left(\tilde{y}_i = y_i \,|\, \tilde{x}_i = x_i\right)
\end{aligned}
$$

Likelihood and log-likelihood

$$p_{\alpha,\beta}(x) := \mathrm{lgf}(\beta^T x + \alpha)$$

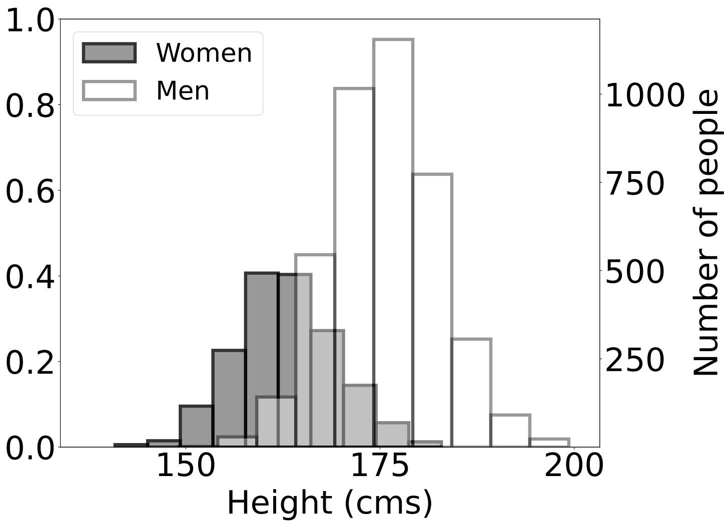
$$\begin{aligned}\mathcal{L}_{XY}(\alpha,\beta) &= \prod_{i=1}^{n} \mathrm{P}\left(\tilde{y}_i = y_i \,|\, \tilde{x}_i = x_i\right) \\ &= \prod_{\{i: y_i = 0\}} \left(1 - p_{\alpha,\beta}(x_i)\right) \prod_{\{l: y_l = 1\}} p_{\alpha,\beta}(x_l)\end{aligned}$$

$$\log \mathcal{L}_{XY}(\alpha,\beta) = \sum_{\{i: y_i = 0\}} \log\left(1 - p_{\alpha,\beta}(x_i)\right) + \sum_{\{l: y_l = 1\}} \log p_{\alpha,\beta}(x_l)$$

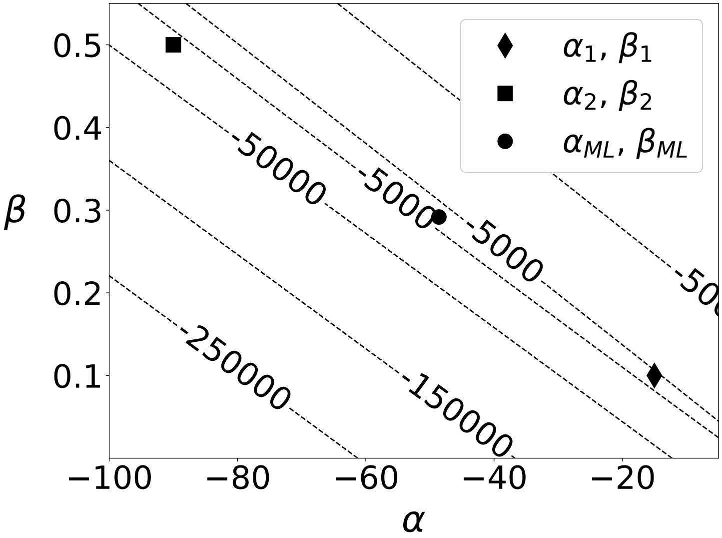
No closed-form solution, but concave

Maximized via iterative methods (gradient ascent, Newton's method, iterative reweighted least squares)

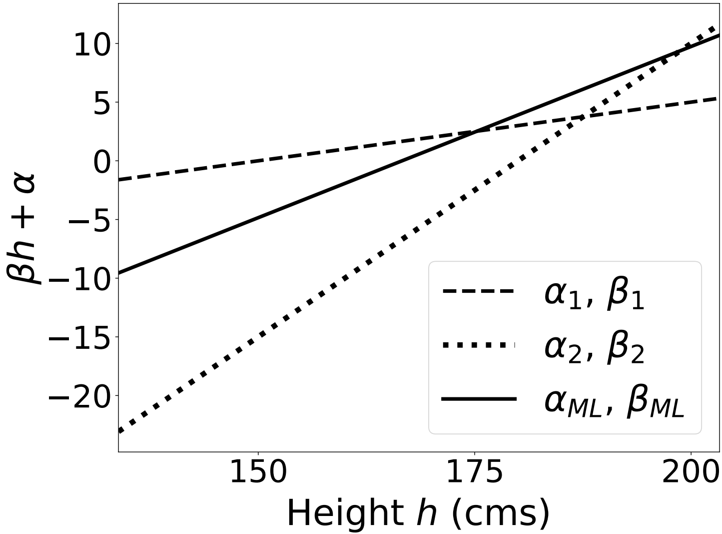
Classification according to height



Log-likelihood

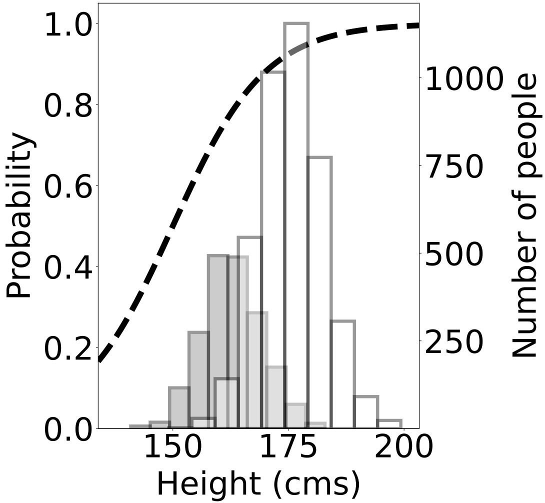


Logits

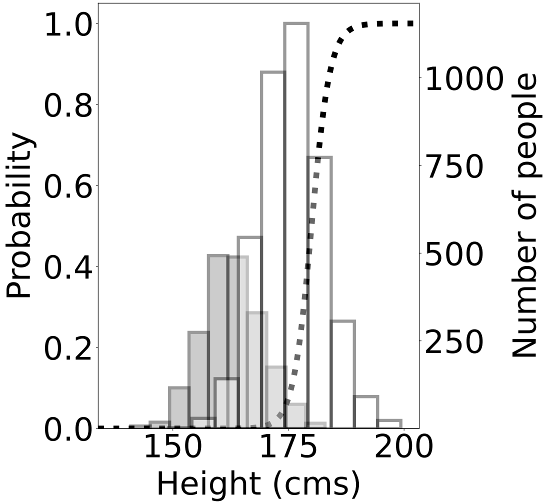


10
5
0
−5
−10
−15
−20
$\beta h + \alpha$
150
175
200
Height h (cms)
α_1, β_1
α_2, β_2
α_{ML}, β_{ML}

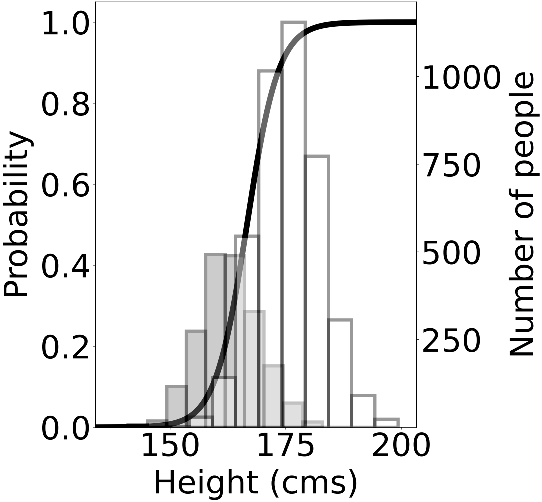
$\alpha_1 := -15,\ \beta_1 := 0.1$



$\alpha_2 := -90,\ \beta_2 := 0.5$



$\alpha_{\mathrm{ML}} := -48.6,\ \beta_{\mathrm{ML}} := 0.29$



Diagnosis of Alzheimer's disease

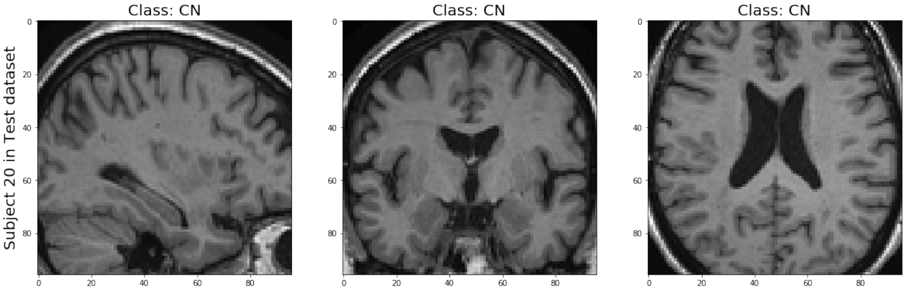
Neurodegenerative disease causing $60 - 70\%$ cases of dementia

Diagnosis via positron-emission tomography is invasive and very costly

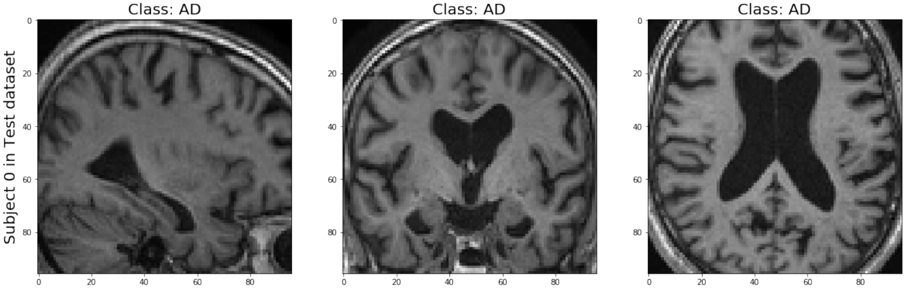
Structural MRI is non-invasive and less costly

Goal: Diagnose Alzheimer's using MRI scans

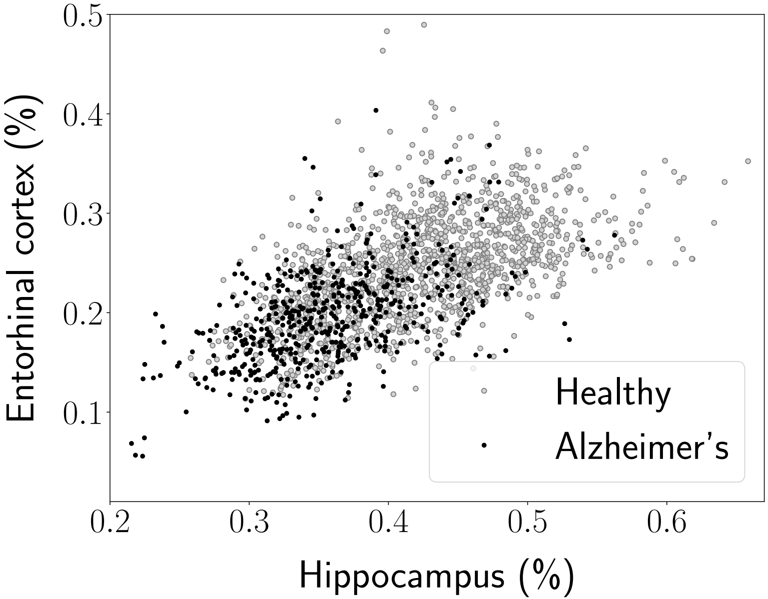
Cognitively-normal patient

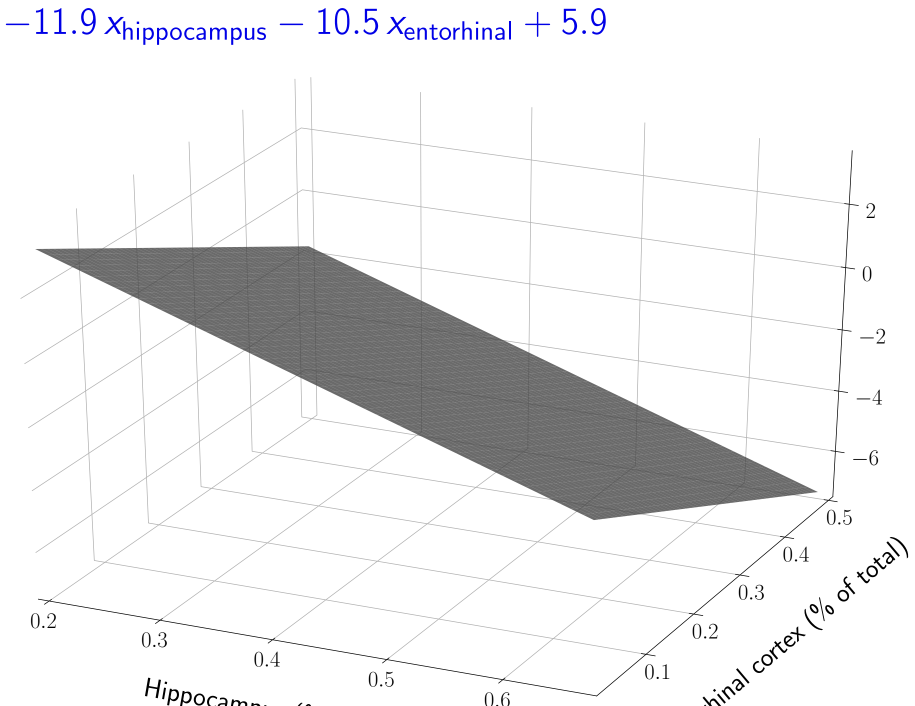


Alzheimer's patient

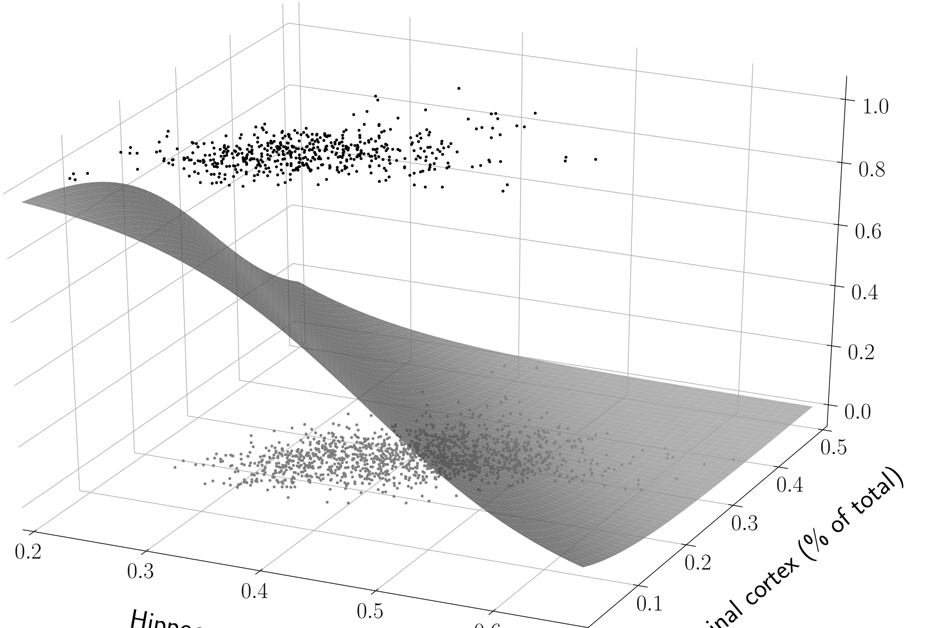


Alzheimer's diagnosis

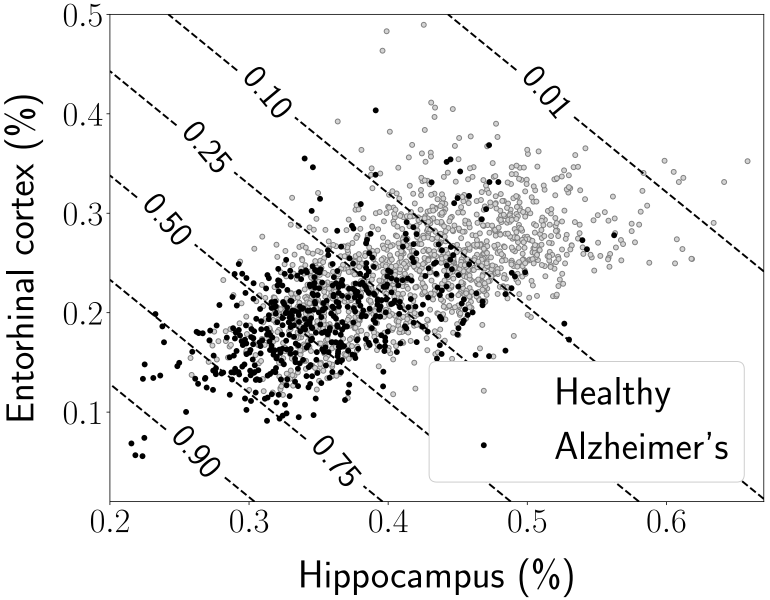



$-11.9\,x_{\text{hippocampus}} - 10.5\,x_{\text{entorhinal}} + 5.9$
2
0
−2
−4
−6
0.2
0.3
0.4
0.5
0.6
Hippocamp
0.1
0.2
0.3
0.4
0.5
hinal cortex (% of total)

$$\mathrm{lgf}\left(-11.9\, x_{\mathrm{hippocampus}} - 10.5\, x_{\mathrm{entorhinal}} + 5.9\right)$$



$\mathrm{lgf}\,(-11.9\,x_{\mathrm{hippocampus}} - 10.5\,x_{\mathrm{entorhinal}} + 5.9)$



Regularization

Feature collinearity produces noisy coefficients and overfitting (as in linear regression)

Solution: Regularization

Regularized cost function:

$$-\log \mathcal{L}_{XY}(\alpha, \beta) + \lambda \, ||\beta||_2^2$$

where λ is a regularization parameter

What have we learned?

How logistic regression works:

- Link function maps linear functions of features (logits) to probability estimates
- Parameters are obtained by maximizing the likelihood