Gaussian Discriminant Analysis

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

Goals

Explain how to use Gaussian mixture models for classification

Motivation: Diagnosis of Alzheimer's disease

Diagnosis of Alzheimer's disease

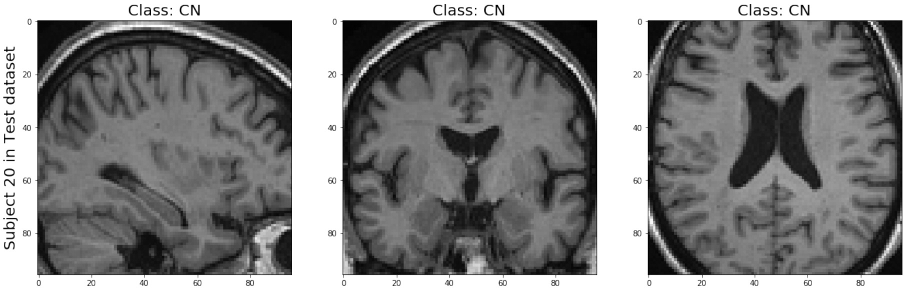
Neurodegenerative disease causing $60-70\%$ cases of dementia

Diagnosis via positron-emission tomography is invasive and very costly

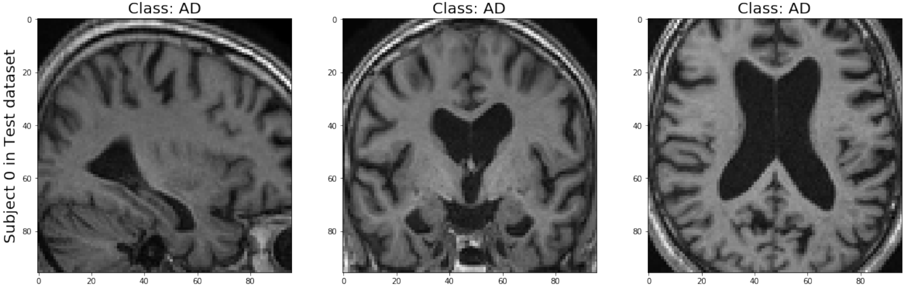
Structural MRI is non-invasive and less costly

Goal: Diagnose Alzheimer's using MRI scans

Cognitively-normal patient



Alzheimer's patient



Classification

Data: (x_1, y_1), (x_2, y_2), ..., (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 class as random variable $\tilde{y}$

For new data vector x:

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

Is classification easy?

Curse of dimensionality

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

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

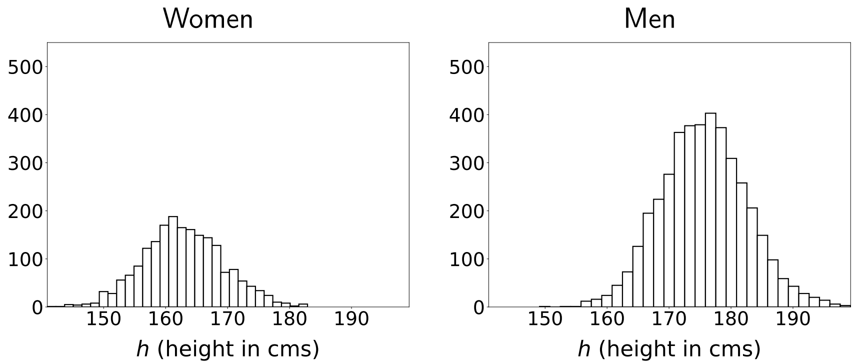
Possible solution: Assume conditional independence of features given class (Naive Bayes)

Alternative: Use parametric model

Parametric mixture model

Assumption: Distribution of features given class y is parametric, with parameters that depend on y

Classification according to height



Classification according to height

Height: Continuous random variable $\tilde{h}$

Sex: Discrete random variable $\tilde{s}$

Assumption: Conditional distribution of $\tilde{h}$ given $\tilde{s} = s$ is Gaussian with parameters that depend on s

Gaussian random vector

A Gaussian random vector $\tilde{x}$ is a random vector with joint pdf

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

where $\mu \in \mathbb{R}^d$ is the mean and $\Sigma \in \mathbb{R}^{d \times d}$ the covariance matrix

$\Sigma \in \mathbb{R}^{d \times d}$ is symmetric and positive definite (positive eigenvalues)

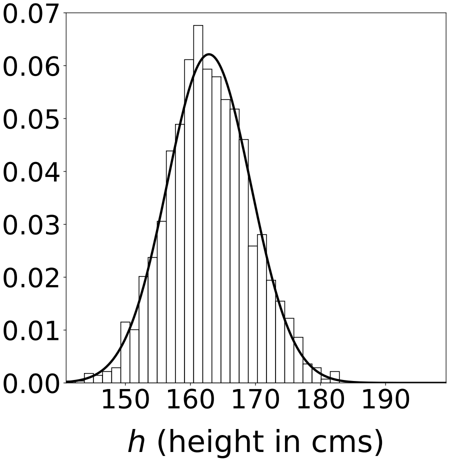
Maximum likelihood estimates

$$\mu_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\Sigma_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\mathrm{ML}})(x_i - \mu_{\mathrm{ML}})^T$$

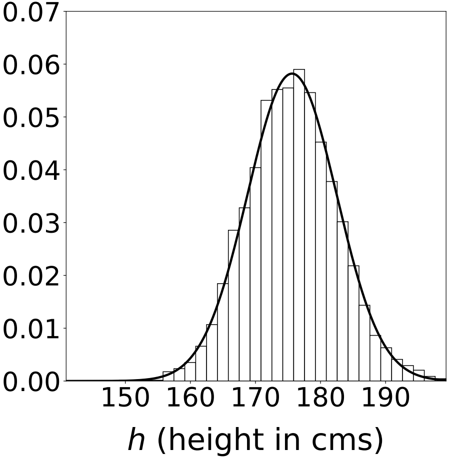
Conditional distribution of $\tilde{h}$ given $\tilde{s}$ = woman

Gaussian with $\mu_{\text{women}} = 163$ cm and $\sigma_{\text{women}} = 6.4$ cm



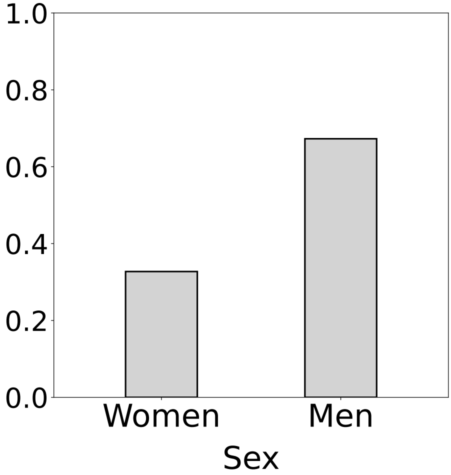
Conditional distribution of $\tilde{h}$ given $\tilde{s}$ = man

Gaussian with $\mu_{\text{men}} = 176$ cm and $\sigma_{\text{men}} = 6.9$ cm



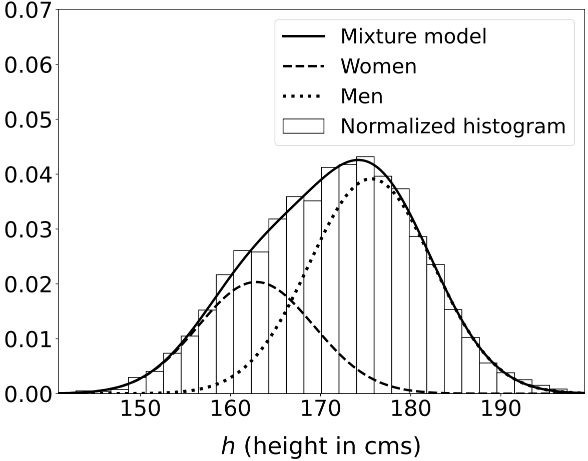
Marginal distribution of $\tilde{s}$

1,986 women and 4,082 men



Gaussian mixture model

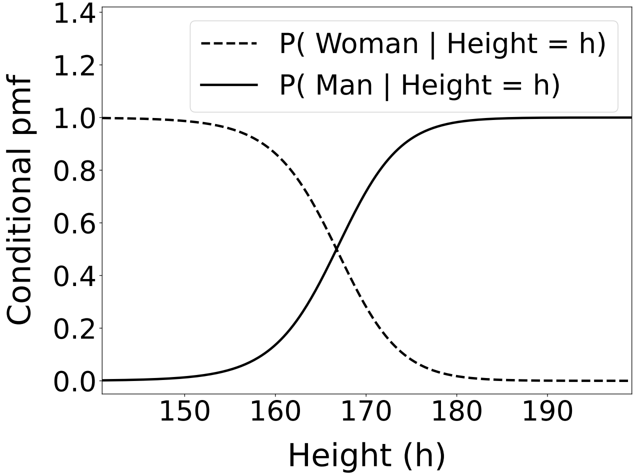
$$f_{\tilde{h}}(h) = p_{\tilde{s}}(\text{woman})\, f_{\tilde{h}\,|\,\tilde{s}}(h\,|\,\text{woman}) + p_{\tilde{s}}(\text{man})\, f_{\tilde{h}\,|\,\tilde{s}}(h\,|\,\text{man})$$



Conditional distribution of $\tilde{s}$ given $\tilde{h}$?

$$
\begin{aligned}
& p_{\tilde{s}\,|\,\tilde{h}}\left(\text{woman}\,|\,h\right) \\
&= \frac{p_{\tilde{s}}\left(\text{woman}\right) f_{\tilde{h}\,|\,\tilde{s}}\left(h\,|\,\text{woman}\right)}{f_{\tilde{h}}\left(h\right)} \\
&= \frac{p_{\tilde{s}}\left(\text{woman}\right) f_{\tilde{h}\,|\,\tilde{s}}\left(h\,|\,\text{woman}\right)}{p_{\tilde{s}}\left(\text{woman}\right) f_{\tilde{h}\,|\,\tilde{s}}\left(h\,|\,\text{woman}\right) + p_{\tilde{s}}\left(\text{man}\right) f_{\tilde{h}\,|\,\tilde{s}}\left(h\,|\,\text{man}\right)} \\
&= \frac{\frac{p_{\tilde{s}}(\text{woman})}{\sqrt{2\pi}\sigma_{\textbf{women}}} \exp\left(-\frac{1}{2}\left(\frac{h-\mu_{\textbf{women}}}{\sigma_{\textbf{women}}}\right)^2\right)}{\frac{p_{\tilde{s}}(\text{woman})}{\sqrt{2\pi}\sigma_{\textbf{women}}} \exp\left(-\frac{1}{2}\left(\frac{h-\mu_{\textbf{women}}}{\sigma_{\textbf{women}}}\right)^2\right) + \frac{p_{\tilde{s}}(\text{man})}{\sqrt{2\pi}\sigma_{\textbf{men}}} \exp\left(-\frac{1}{2}\left(\frac{h-\mu_{\textbf{men}}}{\sigma_{\textbf{men}}}\right)^2\right)} \\
&= \frac{1}{1 + \frac{p_{\tilde{s}}(\text{man})}{p_{\tilde{s}}(\text{woman})} \frac{\sigma_{\textbf{women}}}{\sigma_{\textbf{men}}} \exp\left(\frac{1}{2}\left(\frac{h-\mu_{\textbf{women}}}{\sigma_{\textbf{women}}}\right)^2 - \frac{1}{2}\left(\frac{h-\mu_{\textbf{men}}}{\sigma_{\textbf{men}}}\right)^2\right)} \\
&= \frac{1}{1 + 0.7 \exp\left(0.0017 h^2 - 0.28 h\right)}
\end{aligned}
$$

Conditional pmf of $\tilde{s}$ given $\tilde{h}$



Gaussian discriminant analysis

Idea: Use Gaussian mixture model for classification

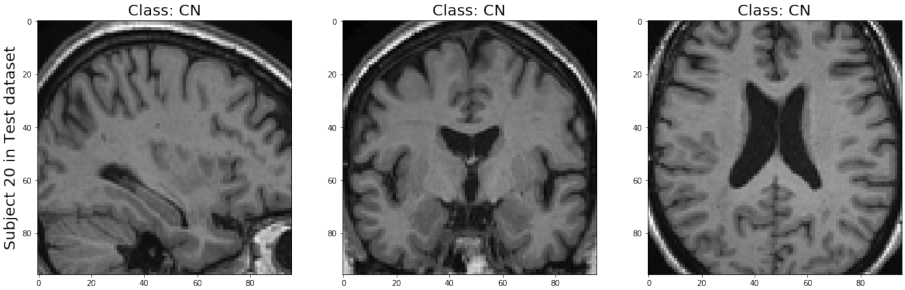
1. $f_{\tilde{x} \mid \tilde{y}}$: For class y, fit Gaussian to training examples with label y to obtain μ_y and Σ_y
2. $p_{\tilde{y}}$: Set $p_{\tilde{y}}(y)$ to fraction of examples in class y
3. Classify test data based on $p_{\tilde{y} \mid \tilde{x}}(\cdot \mid x_{\text{test}})$

Number of parameters scales quadratically with number of features

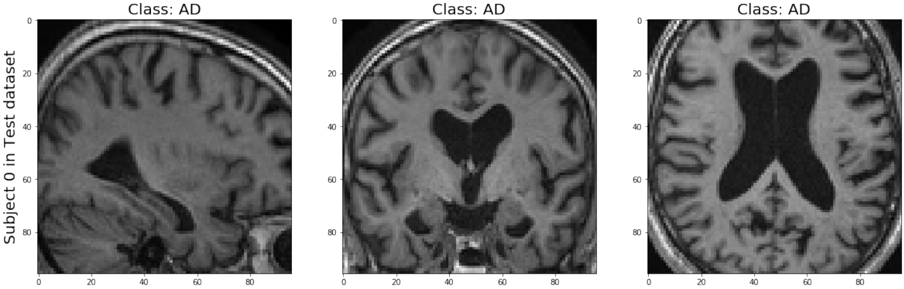
Diagnosis of Alzheimer's disease

Goal: Diagnose Alzheimer's using MRI scans

Cognitively-normal patient

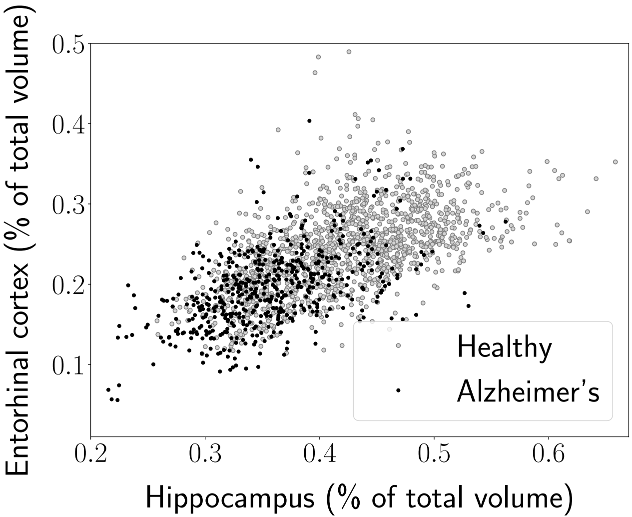


Alzheimer's patient

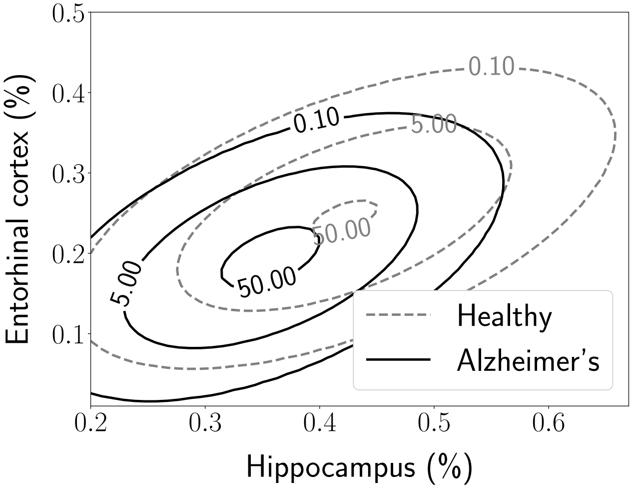


Training data

Alzheimer's Disease Neuroimaging Initiative



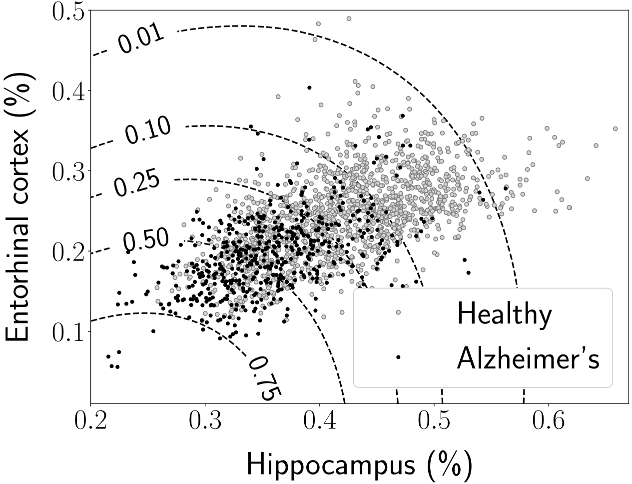
$f_{\tilde{x} \mid \tilde{y}}$



Classification

$$
\begin{aligned}
&\arg \max_{y \in \{1,2,\dots,c\}} p_{\tilde{y} \mid \tilde{x}}(y \mid x) \\
&= \arg \max_{y \in \{1,2,\dots,c\}} \frac{p_{\tilde{y}}(y) f_{\tilde{x} \mid \tilde{y}}(x \mid y)}{f_{\tilde{x}}(x)} \\
&= \arg \max_{y \in \{1,2,\dots,c\}} \frac{p_{\tilde{y}}(y) f_{\tilde{x} \mid \tilde{y}}(x \mid y)}{\sum_{k \in \{1,2,\dots,c\}} p_{\tilde{y}}(k) f_{\tilde{x} \mid \tilde{y}}(x \mid k)} \\
&= \arg \max_{y \in \{1,2,\dots,c\}} \frac{\frac{p_{\tilde{y}}(y)}{\sqrt{(2\pi)^d \left|\Sigma_y\right|}} \exp\left(-\frac{1}{2}\left(x - \mu_y\right)^T \Sigma_y^{-1} \left(x - \mu_y\right)\right)}{\sum_{k \in \{1,2,\dots,c\}} \frac{p_{\tilde{y}}(k)}{\sqrt{(2\pi)^d \left|\Sigma_k\right|}} \exp\left(-\frac{1}{2}\left(x - \mu_k\right)^T \Sigma_k^{-1} \left(x - \mu_k\right)\right)}
\end{aligned}
$$

$p_{\tilde{y} \mid \tilde{x}}$



Training error

We diagnose Alzheimer's if $p_{\tilde{y} \mid \tilde{x}}(1 \mid x) > 0.5$

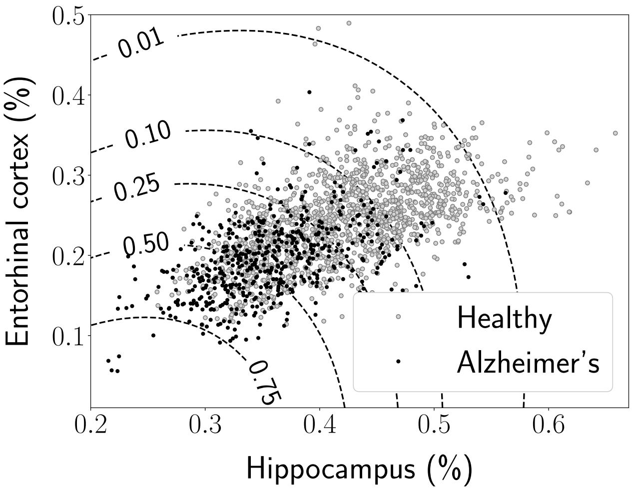
Error rate on training data: 24.2%

Fraction of Alzheimer's patients: 27.1%

Is this what we care about?

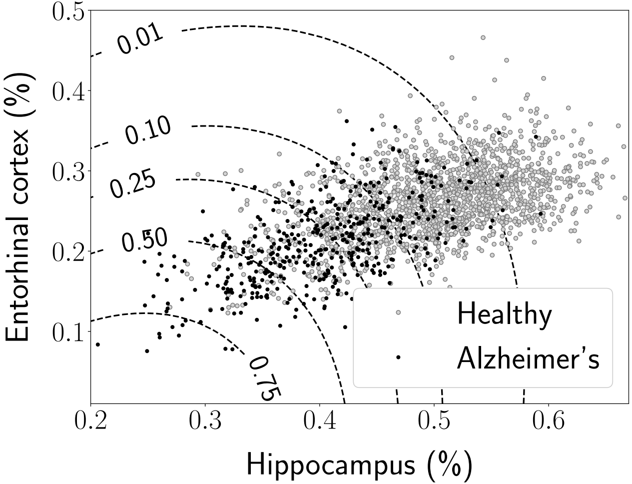
Training data

Alzheimer's Disease Neuroimaging Initiative



Test data

National Alzheimer's Coordinating Center



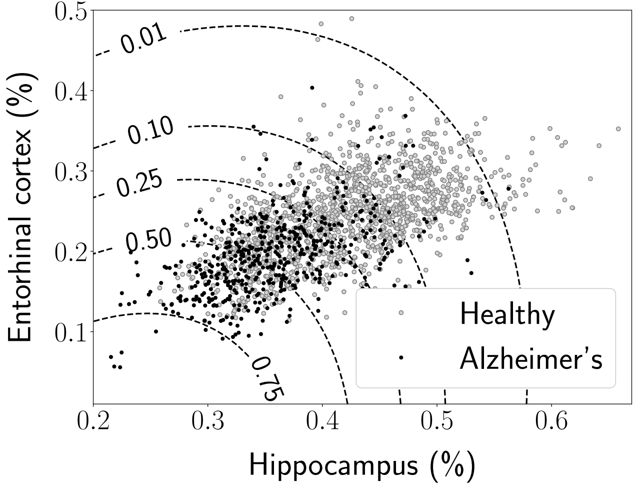
Test error

We diagnose Alzheimer's if $p_{\tilde{y} \mid \tilde{x}}(1 \mid x_{\text{test}}) > 0.5$

Error rate on test data: 18.5%

Fraction of Alzheimer's patients: 21.6%

Decision boundary?



Decision boundary

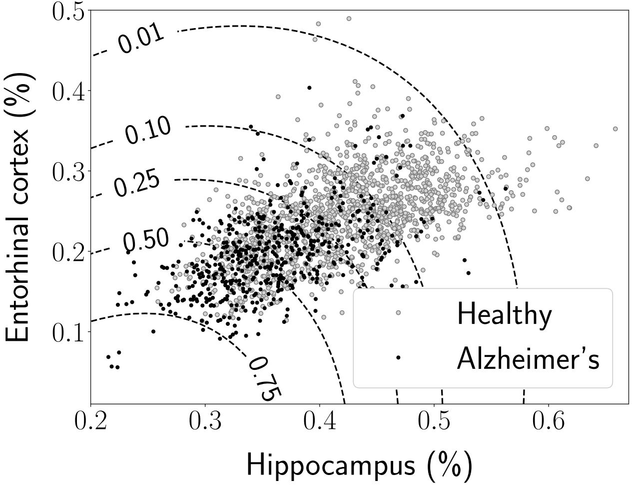
$$p_{\tilde{y}\,|\,\tilde{x}}(y\,|\,x) = \frac{\frac{p_{\tilde{y}}(y)}{\sqrt{(2\pi)^d\left|\Sigma_y\right|}} \exp\left(-\frac{1}{2}\left(x-\mu_y\right)^T \Sigma_y^{-1}\left(x-\mu_y\right)\right)}{\sum_{k\in\{1,2,\ldots,c\}} \frac{p_{\tilde{y}}(k)}{\sqrt{(2\pi)^d\left|\Sigma_k\right|}} \exp\left(-\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right)\right)}$$

Decision boundary: $1 = \dfrac{p_{\tilde{y}\,|\,\tilde{x}}(a\,|\,x)}{p_{\tilde{y}\,|\,\tilde{x}}(b\,|\,x)}$

$$1 = \frac{p_{\tilde{y}}(a)\sqrt{|\Sigma_b|}}{p_{\tilde{y}}(b)\sqrt{|\Sigma_a|}} \exp\left(\frac{1}{2}\left(x-\mu_b\right)^T \Sigma_b^{-1}\left(x-\mu_b\right) - \frac{1}{2}\left(x-\mu_a\right)^T \Sigma_a^{-1}\left(x-\mu_a\right)\right)$$

$$0 = \frac{1}{2}\left(x-\mu_b\right)^T \Sigma_b^{-1}\left(x-\mu_b\right) - \frac{1}{2}\left(x-\mu_a\right)^T \Sigma_a^{-1}\left(x-\mu_a\right) + \log \frac{p_{\tilde{y}}(a)\sqrt{|\Sigma_b|}}{p_{\tilde{y}}(b)\sqrt{|\Sigma_a|}}$$

Quadratic discriminant analysis



Decision boundary

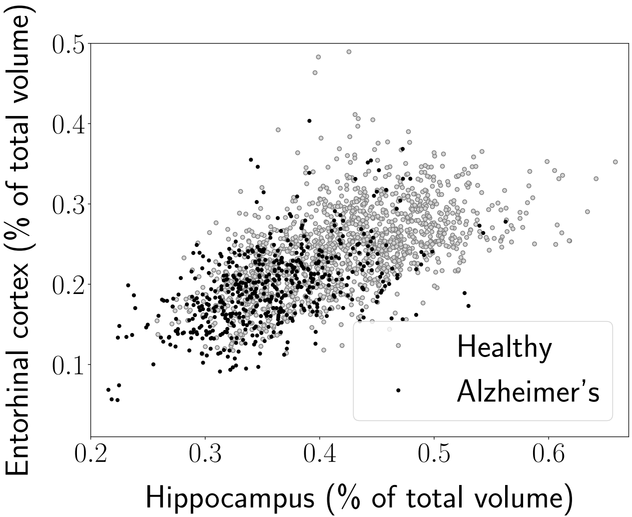
How can we get a linear decision boundary?

$$
\begin{aligned}
0 &= \frac{1}{2}\left(x-\mu_b\right)^T \Sigma_b^{-1}\left(x-\mu_b\right) - \frac{1}{2}\left(x-\mu_a\right)^T \Sigma_a^{-1}\left(x-\mu_a\right) + \log \frac{p_{\tilde{y}}(a)\sqrt{|\Sigma_b|}}{p_{\tilde{y}}(b)\sqrt{|\Sigma_a|}} \\
&= \frac{1}{2}x^T\Sigma_b^{-1}x - \frac{1}{2}x^T\Sigma_a^{-1}x + \text{affine function of } x
\end{aligned}
$$

Set $\Sigma_a = \Sigma_b = \Sigma$

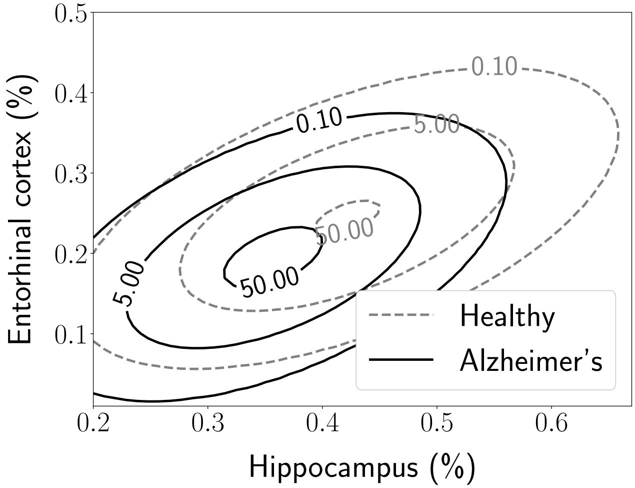
Training data

Alzheimer's Disease Neuroimaging Initiative



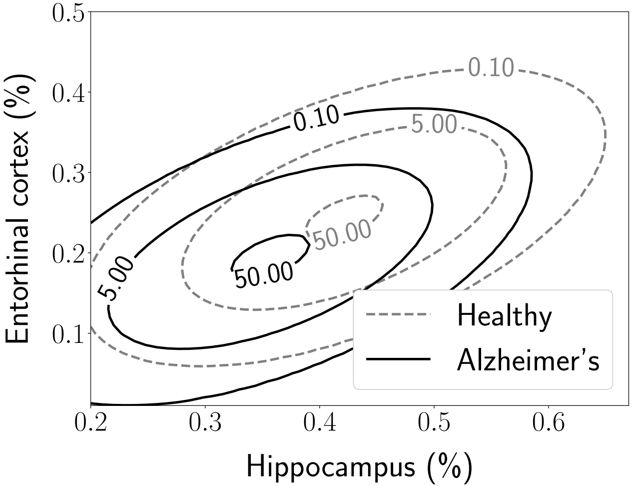
Quadratic discriminant analysis

We fit Σ_a and Σ_b separately

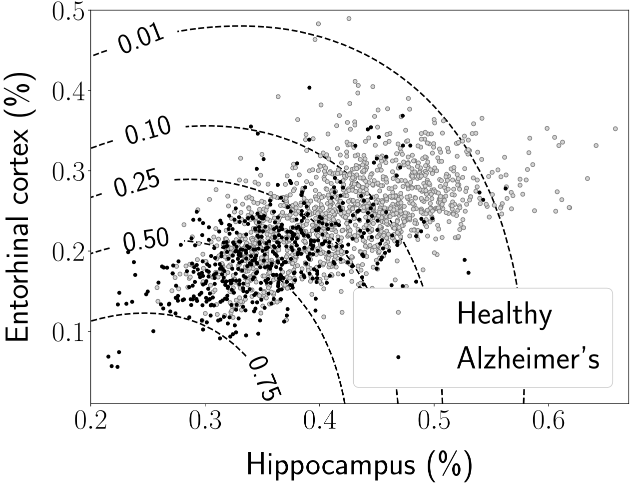


Linear discriminant analysis

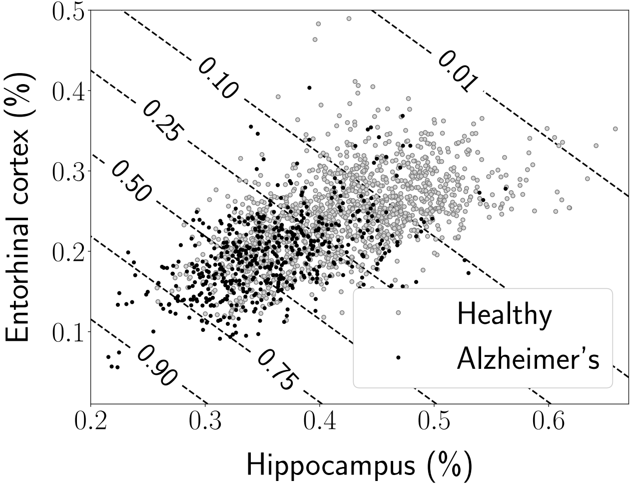
We fit $\Sigma_a = \Sigma_b = \Sigma$



Quadratic discriminant analysis



Linear discriminant analysis



Evaluation

Training error rate: 24.1% (QDA: 24.2%)

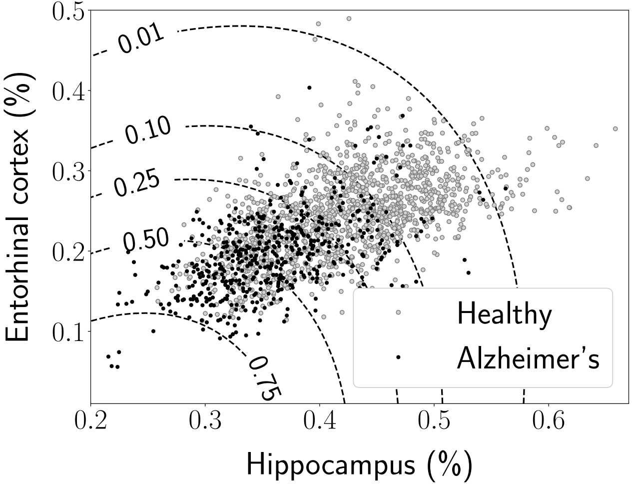
Fraction of Alzheimer's patients: 27.1%

Test error rate: 18.5% (QDA: 18.5%)

Fraction of Alzheimer's patients: 21.6%

How would you improve the model?

How would you improve the model?



What have we learned?

How to use Gaussian mixture models to perform classification