Day 6 - Logistic Regression (continued)

Agenda:

- Classification and Logistic Regression
    - Training binary classifiers
    - Evaluating classifiers
    - Training multiclass classifiers

More thoughts on square capital example and whether to approach problem as regression or classification

Parametric Approach: Choose a model for f with unknown parameters. Estimate the parameters.

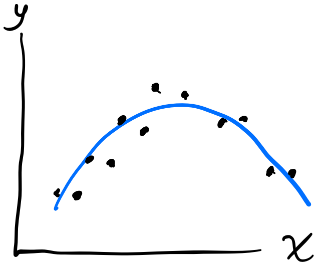
Parametric Regression: predict a continuous value

Model $f_\theta : \mathbb{R}^d \to \mathbb{R}$

$y = f_\theta(x) + \text{noise}$

Given: $\{(x^{(i)}, y_i)\}_{i=1\dots n}$

Find: θ



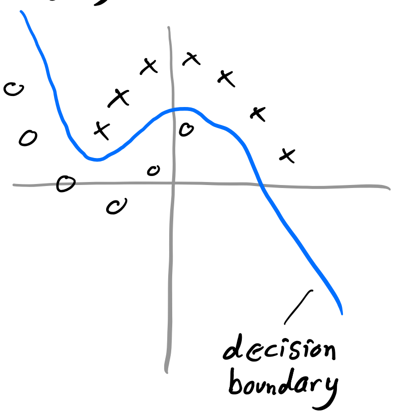
Parametric Classification: predict membership in a category

Let $f_\theta : \mathbb{R}^d \to \left\{ \begin{matrix} \text{cat } 1 \\ \vdots \\ \text{cat } m \end{matrix} \right\}$

$y = f_\theta(x) + \text{noise}$

Given: $\{(x^{(i)}, y_i)\}_{i=1\dots n}$

Find: θ



Approach for estimating θ:

Select a model for f w/ parameters θ

&

minimize the loss between training labels and predictions on training data

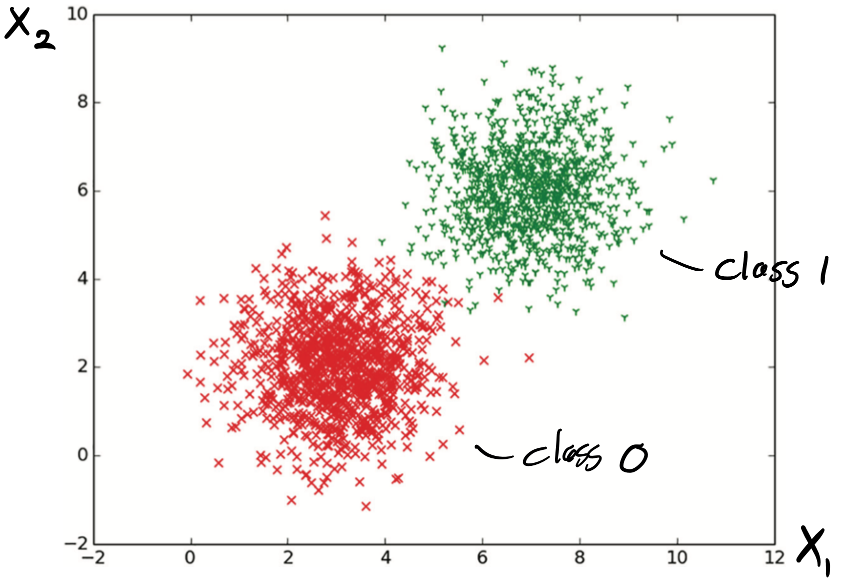
$$\min_{\theta} \sum_{i=1}^{n} \underset{\substack{\text{loss}\\ \text{function}}}{L}(y_i, f_\theta(x^{(i)}))$$

Example: linear regression $L(y, \overset{\substack{\text{prediction}\\ \text{of } y}}{\hat{y}}) = |y-\hat{y}|^2$

"square loss" or ℓ^2 loss

Binary Classification in 2D with logistic regression

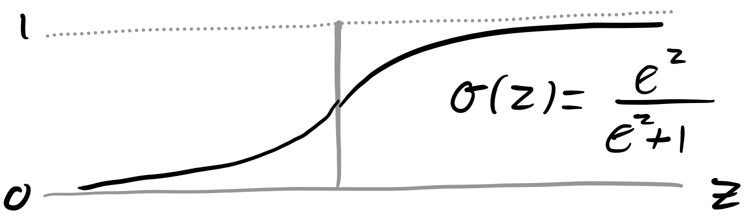
Training Data $\{\overset{\mathbb{R}^2}{x^{(i)}}, \overset{\mathbb{R}}{y_i}\}_{i=1\cdots n}$ $\quad y_i = \begin{cases} 1 & \text{if class } 1 \\ 0 & \text{if class } 0 \end{cases}$



Model

$$y = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2) = \hat{y}(x;\theta)$$

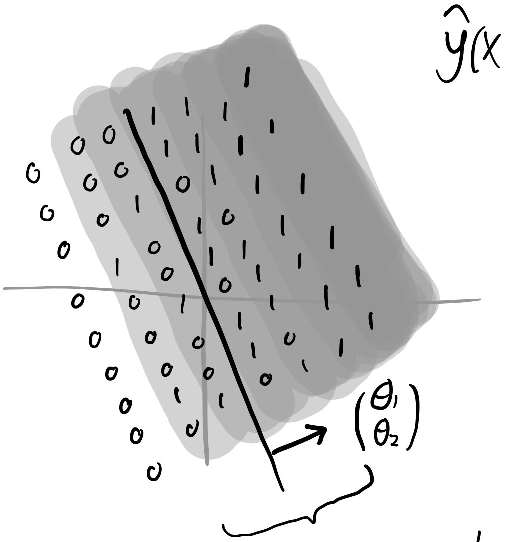
where



$$\sigma(z) = \frac{e^z}{e^z+1}$$

"sigmoid"

"logistic function"



$\hat{y}(x;\theta)$ – confidence that x belongs to class 1

width of region of uncertainty $\simeq \frac{1}{\left\|\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}\right\|}$

Solve

$$\min_{\theta} \sum_{i=1}^{n} L\left(y_i,\ \hat{y}(x^{(i)};\theta)\right) \text{ for } \hat{\theta}$$

Predict:

For new sample x, predict

$$\begin{cases} \text{class } 1 & \text{if } \hat{y} \geq 1/2 \\ \text{class } 0 & \text{if } \hat{y} < 1/2 \end{cases}$$

What loss function should you use?

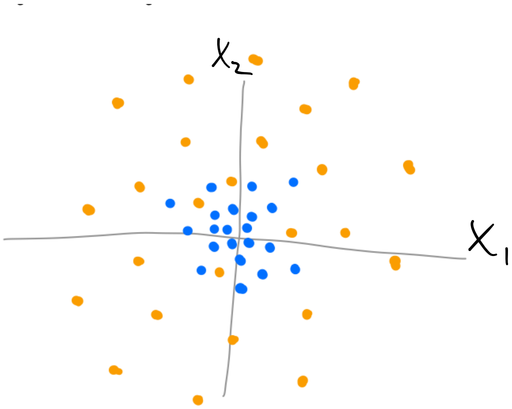
One choice - log loss

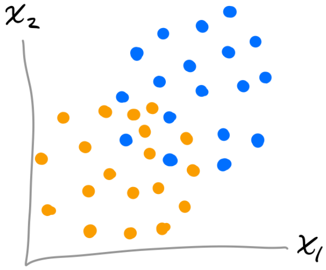
$$L(\underset{\text{binary}}{y}, \underset{\text{continuous}}{\hat{y}}) = \begin{cases} -\log(\hat{y}) & \text{if } y=1 \\ -\log(1-\hat{y}) & \text{if } y=0 \end{cases}$$

$$= -y\log\hat{y} - (1-y)\log(1-\hat{y})$$

Question:

Consider a binary classification problem. What property of the data would lead logistic regression to learn a good classifier?



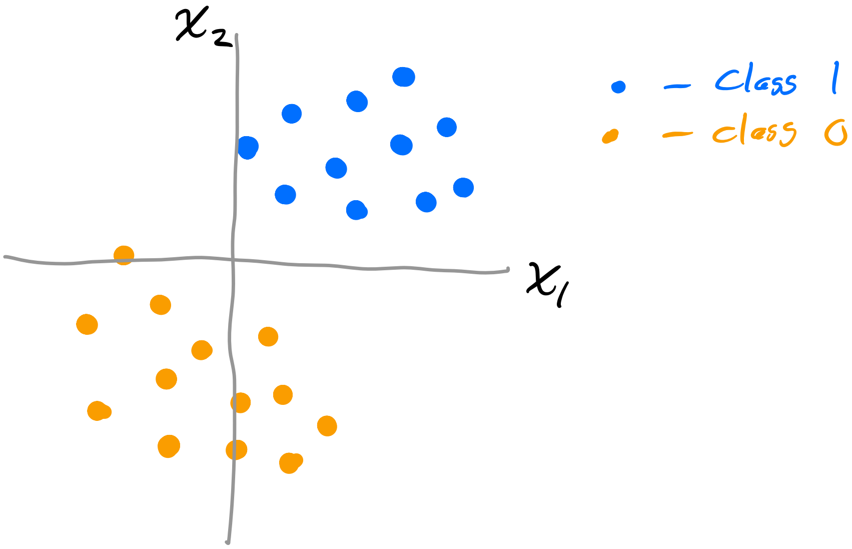


Question: With the following data, does a minimizer to the optimization problem even exist?

$$\min_{\theta} \sum_{i=1}^{n} L\left(y_i, \hat{y}(x^{(i)};\theta)\right)$$

$$L(y,\hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

$$\hat{y} = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$



Evaluating Classifiers

		Prediction	
		+	−
Truth	+	True Positive (TP)	False Negative (FN)
	−	False Positive (FP)	True Negative (TN)

Accuracy: $\dfrac{TP + TN}{\text{total}}$

Activity: Someone invents a test for a rare disease that affects 0.1% of the population. The test has accuracy 99.9%. Are you convinced this is a good test?

Precision : $\frac{TP}{TP+FP}$ What fraction of predicted positives are real?

Recall : $\frac{TP}{TP+FN}$ What fraction of positives are correctly predicted?

Want high precision & high recall

Activity: You are building a binary classifier that detects whether a pedestrian is crossing the sidewalk within 30 feet of a self driving car. If the detection is positive, the car puts on the breaks. Would you rather have good precision and great recall or good recall and great precision?

There is a trade off between True Positives and False Positives, and between True Negatives and False Negatives

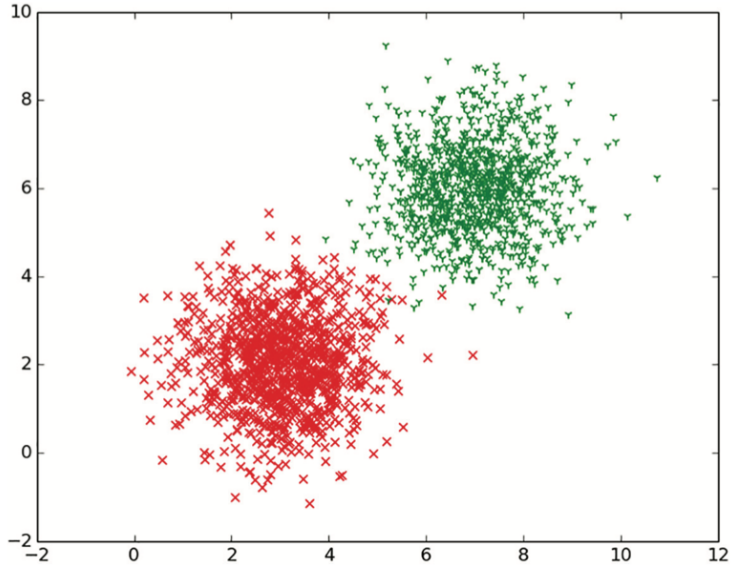
Training Data: $\{ \overset{\mathbb{R}^2}{x^{(i)}}, \overset{\mathbb{R}}{y_i} \}_{i=1\ldots n}$ $y_i = \begin{cases} 1 & \text{if class } 1 \\ 0 & \text{if class } 0 \end{cases}$

Model: $y = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2) = \hat{y}(x;\theta)$

Predict: For new sample x, predict

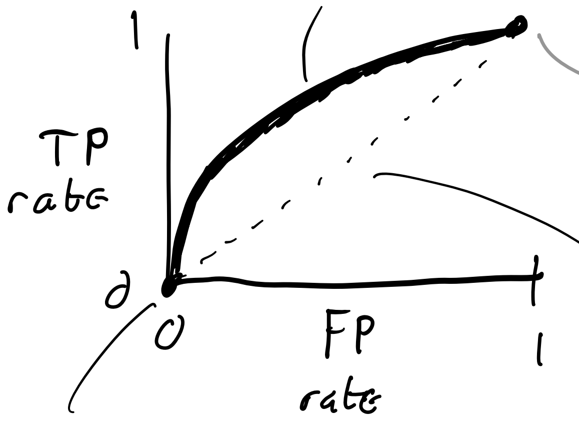
$$\begin{cases} \text{class } 1 & \text{if } \hat{y} \geq 1/2 \\ \text{class } 0 & \text{if } \hat{y} < 1/2 \end{cases}$$

could choose any value



Receiver Operating Characteristic Curves

Each point is a classifier w/ different threshold



declare everything positive

random classifier

declare everything negative

Comparing classifiers and Area-Under-Curve (AUC)

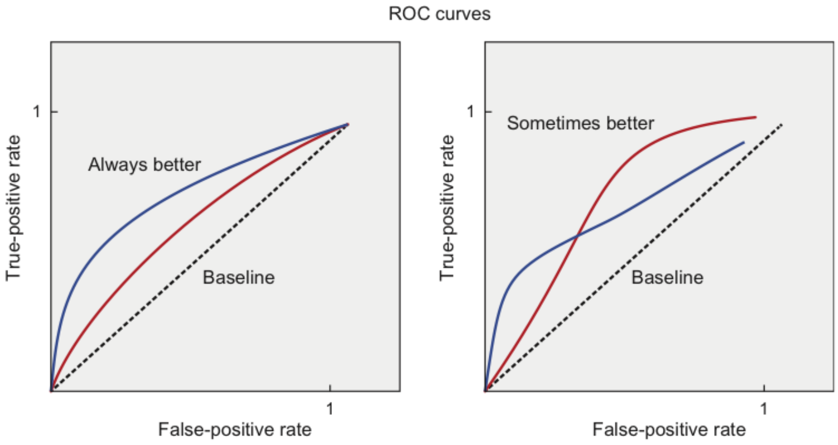
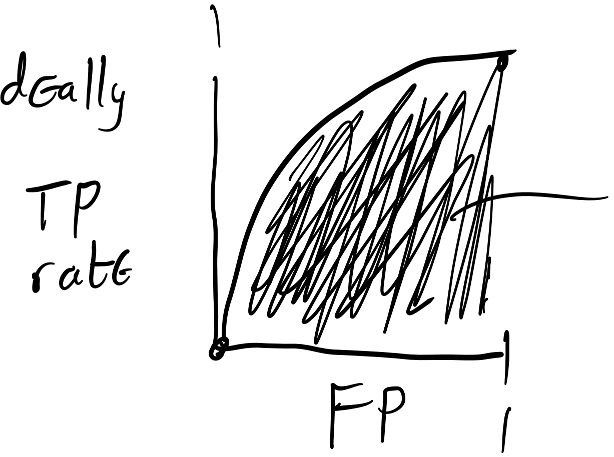


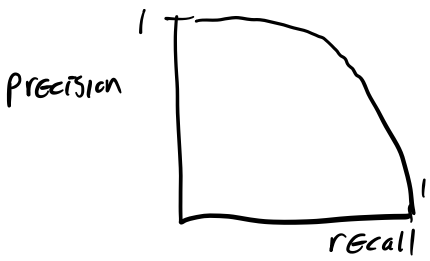
Figure 5.6 The principled way to compare algorithms is to examine their ROC curves. When the true-positive rate is greater than the false-positive rate in every situation, it's straightforward to declare that one algorithm is dominant in terms of its performance. If the true-positive rate is less than the false-positive rate, the plot dips below the baseline shown by the dotted line.

Ideally



want AUC = 1

Also common to plot precision-recall curves



Multiclass Classification

Consider a 3 class classification problem (w/ no bias term)

Training Data: $\{(x^{(i)}, y^{(i)})\}_{i=1\cdots n}$

$$x^{(i)} \in \mathbb{R}^d \quad \text{for all } i$$

$$y^{(i)} \in \mathbb{R}^3, \qquad y^{(i)} = \begin{cases} \begin{pmatrix}1\\0\\0\end{pmatrix} & \text{if } y^{(i)} \text{ of class } 1 \\ \begin{pmatrix}0\\1\\0\end{pmatrix} & \text{if } y^{(i)} \text{ of class } 2 \\ \begin{pmatrix}0\\0\\1\end{pmatrix} & \text{if } y^{(i)} \text{ of class } 3 \end{cases}$$

Model: $y = \text{Softmax}(z_1, z_2, z_3)$

$$\text{logits}\left\{ \begin{aligned} z_1 &= \theta_1 \cdot x \\ z_2 &= \theta_2 \cdot x \\ z_3 &= \theta_3 \cdot x \end{aligned} \right. \qquad \text{w/} \quad \begin{aligned} \theta_1 &\in \mathbb{R}^d \\ \theta_2 &\in \mathbb{R}^d \\ \theta_3 &\in \mathbb{R}^d \end{aligned}$$

"logits are linear in parameters of model"

$$\text{softmax}(z_1, z_2, z_3) = \begin{pmatrix} \dfrac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \\[2ex] \dfrac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \\[2ex] \dfrac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}} \end{pmatrix} \text{— adds up to } 1$$

Train:
$$\min_{\theta} \sum_{i=1}^{n} L\left(y^{(i)}, \text{softmax}\left(\theta_1 \cdot x^{(i)}, \theta_2 \cdot x^{(i)}, \theta_3 \cdot x^{(i)}\right)\right)$$

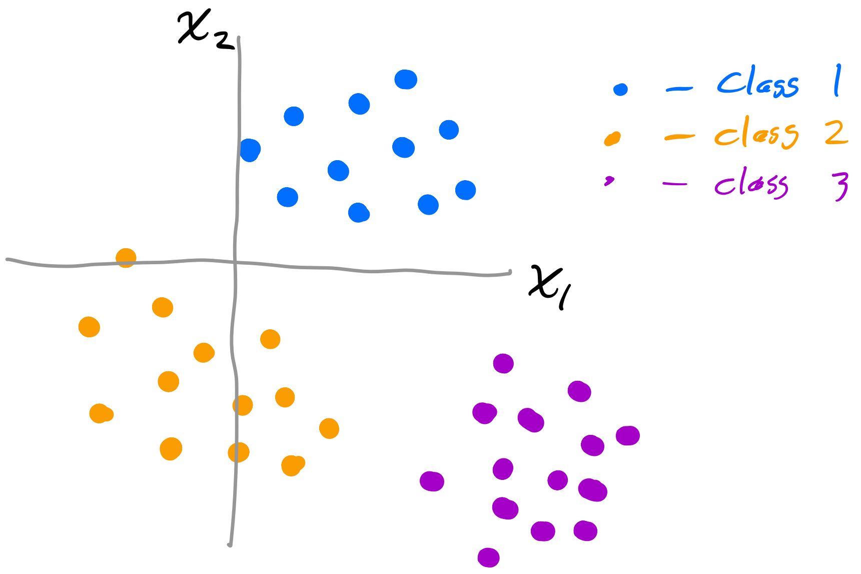
log loss:
$$L(\underset{\overset{|}{\mathbb{R}^3}}{y}, \underset{\overset{|}{\mathbb{R}^3}}{\hat{y}}) = -\sum_{c=1}^{3} y_c \log \hat{y}_c$$

Prediction:

Given x, compute $\hat{y}(x;\hat{\theta}) \in \mathbb{R}^3$

Predict class $\hat{c}$ w/ $\hat{c} = \underset{c}{\operatorname{argmax}}\ \hat{y}_c$

What will decision boundary look like for 3-class classification?



CS 6140: Machine Learning — Fall 2021— Paul Hand

HW 3
Due: Wednesday October 6, 2021 at 2:30 PM Eastern time via Gradescope.

Names: [Put Your Name(s) Here]

You can submit this homework either by yourself or in a group of 2. You may consult any and all resources. Make sure to justify your answers. If you are working alone, you may either write your responses in LaTeX or you may write them by hand and take a photograph of them. If you are working in a group of 2, you must type your responses in LaTeX. You are encouraged to use Overleaf. Create a new project and replace the tex code with the tex file of this document, which you can find on the course website. To share the document with your partner, click Share > Turn on link sharing, and send the link to your partner. When you upload your solutions to Gradescope, make sure to take each problem with the correct page or image.

Question 1. *Linear regression with multivariate responses.*

Consider training data $\{(x^{(i)}, y^{(i)})\}_{i=1\dots n}$, where $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}^k$. Consider a model $y = Ax$, where $A \in \mathbb{R}^{k\times d}$ is unknown. Estimate A by solving least squares linear regression

$$\min_A \sum_{i=1}^{n} \|y^{(i)} - Ax^{(i)}\|^2.$$

(a) Find A in the case of training data $\left\{\left(\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1\\1\\0\end{pmatrix}\right), \left(\begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}0\\1\\1\end{pmatrix}\right), \left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}2\\3\\1\end{pmatrix}\right)\right\}$. You may use a computer to perform linear algebra. Hint: the problem can be simplified by observing that each output dimension can be computed separately from the others. If you use this fact, justify it in your response.

Response:

(b) Consider the case of generic training data. Let Y be the $k \times n$ matrix such that $Y_{ji} = y_j^{(i)}$. Let X be the $n \times d$ matrix where $X_{ij} = x_j^{(i)}$. Provide a formula for the least squares estimate of A. Make sure to check that the matrix dimensions match in any matrix products that appear in your answer. Use the same hint as in part (a).

Response:

(c) Show that any prediction under this learned model is a linear combination of the response values $(y^{(1)}, \dots, y^{(n)})$. That is, for the A in part (b), show that $Ax \in \text{span}(y^{(1)}, \dots, y^{(n)})$ for any x. You may assume that X is rank d.

Response:

Question 2. *Logistic Regression*

Consider training data $\{(x_i, y_i)\}_{i=1\ldots n}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{0,1\}$. Consider the logistic data model $\hat{y} = \sigma(\theta \cdot x)$, where $x \in \mathbb{R}^d$, $\theta \in \mathbb{R}^d$, and σ is the logistic function $\sigma(z) = e^z/(e^z + 1)$.

(a) Show that $\sigma'(z) = \sigma(z)(1 - \sigma(z))$.

Response:

(b) Let $f(\theta) = \sum_{i=1}^{n} -y_i \log \hat{y}_i - (1 - y_i)\log(1 - \hat{y}_i)$, where $\hat{y}_i = \sigma(\theta \cdot x_i)$. Compute $\nabla f(\theta)$. Use the fact in part (a) to simplify your answer.

Response:

(c) If $M = \sum_{i=1}^{n} x_i x_i^t$, show that $z^t M z \geq 0$ for any $z \in \mathbb{R}^d$.

Response:

(d) Using a summation and vector and/or matrix products, write down a formula for the Hessian, H, of f with respect to θ. Show that $z^t H z \geq 0$ for any $z \in \mathbb{R}^d$.

Response: