Day 6 - Linear Regression and Logistic Regression

Agenda:

- Linear Regression
  - Examples
  - Issues to Pay Attention To with Linear Regression
- Classification and Logistic Regression
  - Training classifiers
  - Evaluating classifiers

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

Least Squares Formulation for Linear Regression (for a general model)

Given: $D = \{(X^{(i)}, y_i)\}_{i=1\cdots n}$, $X^{(i)} \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Model: $$y = \underbrace{\theta_1 g_1(X^{(i)}) + \cdots \theta_k g_k(X^{(i)})}_{f_\theta(x)} + \text{Error}$$

Want $$y \approx \underline{X}\,\theta$$

w/ $$\underline{X} = \begin{pmatrix} g_1(X^{(1)}) & g_2(X^{(1)}) \cdots & g_k(X^{(1)}) \\ \vdots & \ddots & \vdots \\ g_n(X^{(n)}) & \cdots & g_k(X^{(n)}) \end{pmatrix} = \begin{pmatrix} - & g(X^{(1)}) & - \\ - & g(X^{(2)}) & - \\ & \vdots & \\ - & g(X^{(n)}) & - \end{pmatrix}$$

Find $$\min_\theta \frac{1}{2}\|y - \underline{X}\theta\|^2$$

where $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}$

Solution given by <u>Normal Equations</u>

$$X^t X \theta = X^t y \quad \Rightarrow \quad \theta = (X^t X)^{-1} X^t y.$$

Examples of setting up and solving linear regression

Find best fit cubic through 1d data

Data $= \{(x_i, y_i)\}_{i=1\cdots n}$ w/ $x_i, y_i \in \mathbb{R}$

Model $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$ + noise

Find

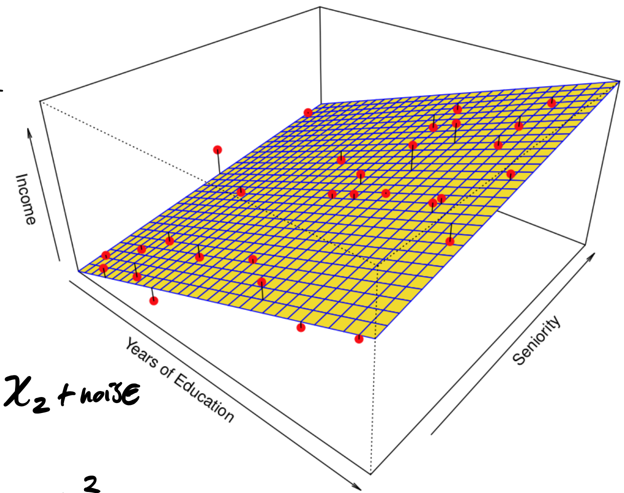
$$\min_\theta \frac{1}{2}\|y - X\theta\|^2$$

where $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\theta = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_3 \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{pmatrix}$

Solution given by Normal Equations

$$X^t X \theta = X^t y \quad \Rightarrow \quad \theta = (X^t X)^{-1} X^t y.$$

Find best fit plane



Data $= \{ (\underbrace{x^{(i)}}_{\mathbb{R}^2}, \underbrace{y_i}_{\mathbb{R}}) \}_{i=1\cdots n}$

Model: $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \text{noise}$

Find

$$\min_{\theta} \frac{1}{2} \| y - X\theta \|^2$$

where $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\theta = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_3 \end{pmatrix}$, $X = \begin{pmatrix} \quad \end{pmatrix}$

Solving and Optimization Problem using Gradient Descent

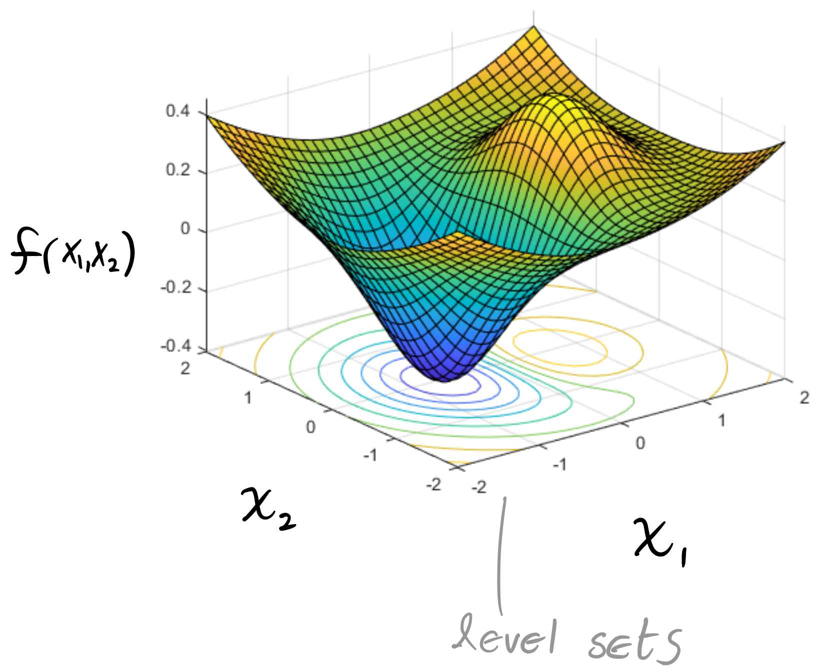
$$\min_{x} \; f(x)$$

Gradient descent: Take successive steps "downhill"

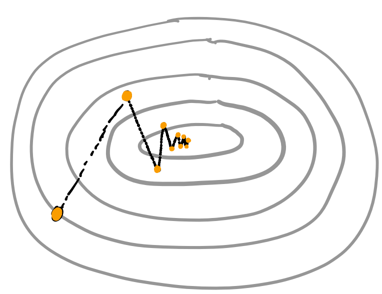
$$x^{i+1} = x^{i} - \alpha \nabla f(x^{i})$$

α: step size, learning rate

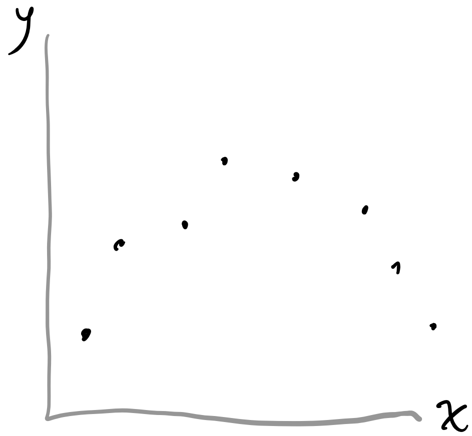
$-\nabla f$ points in direction of steepest descent



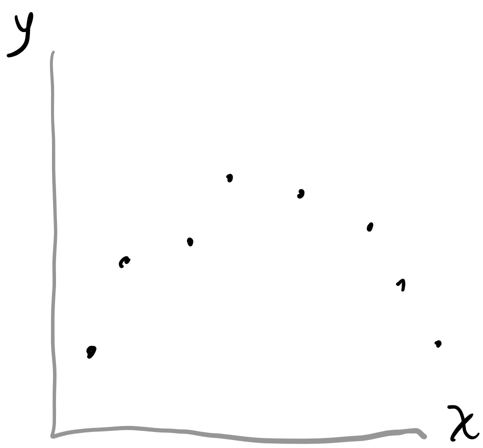
Depiction of gradient descent



Things that can go wrong: Underfitting and Overfitting



Underfitting



Overfitting

Things that can go wrong: numerical instability

Other topics:

What happens when there is fewer data than features?

What happens if there are outliers in the data?

How do you deal with categorical features?

Be careful about whether you want to view your problem as a prediction task

Classification and Logistic Regression

Viewing Regression and Classification as function estimation problems

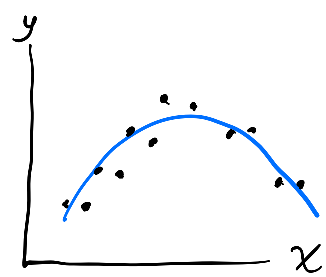
Regression: predict a continuous value

Let $f: \mathbb{R}^d \to \mathbb{R}$

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

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

Find: f



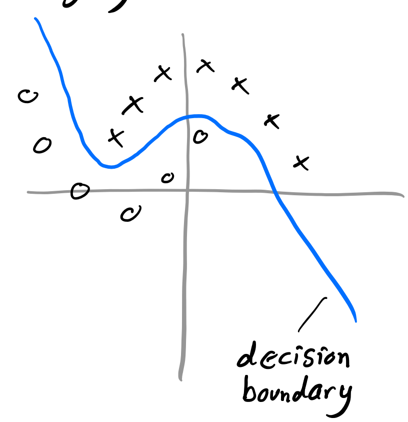
Classification: predict membership in a category

Let $f: \mathbb{R}^d \to \begin{Bmatrix} \text{cat } 1 \\ \vdots \\ \text{cat } m \end{Bmatrix}$

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

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

Find: f



Terminology:

x – input variables, predictors, independent vars, features

y – response, dependent variable, output variable

f – model, predictor, hypothesis

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

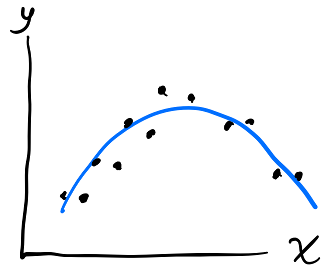
Parametric Regression: predict a continuous value

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

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

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

Find: θ



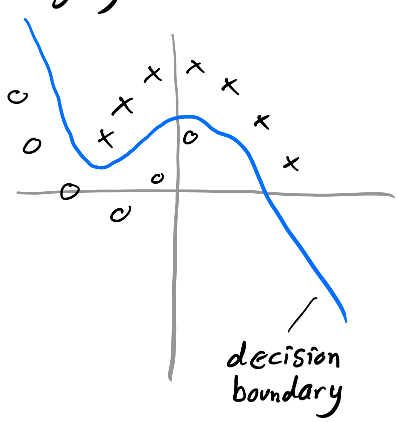
Parametric Classification: predict membership in a category

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

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

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

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

$$\min_{\theta} \sum_{i=1}^{n} L(y_i, f_\theta(x^{(i)}))$$

loss function

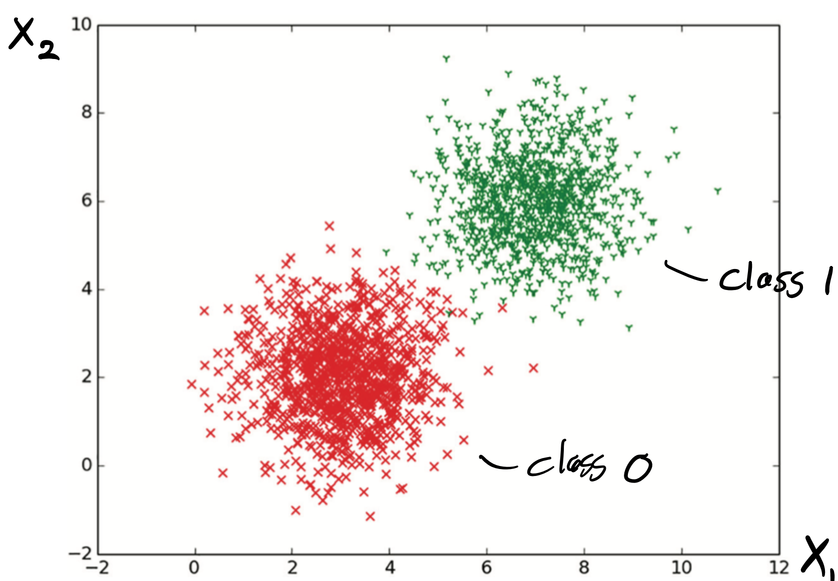
Example: linear regression $L(y, \hat{y}) = |y-\hat{y}|^2$

"square loss" or ℓ^2 loss

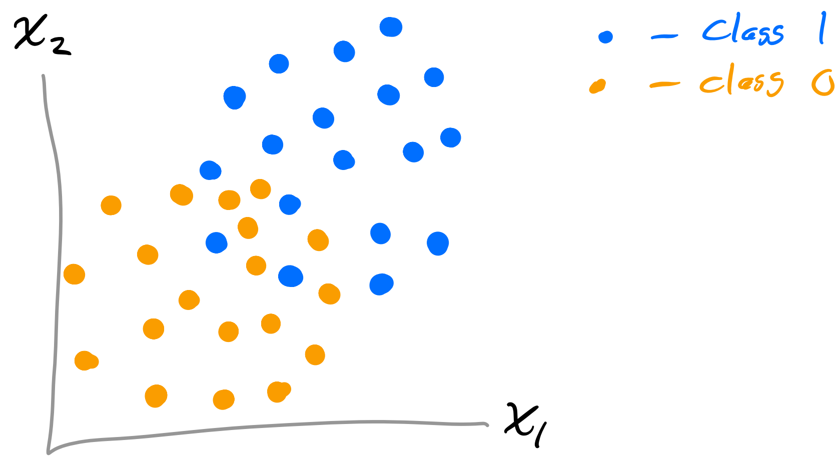
Binary Classification in 2D with logistic regression

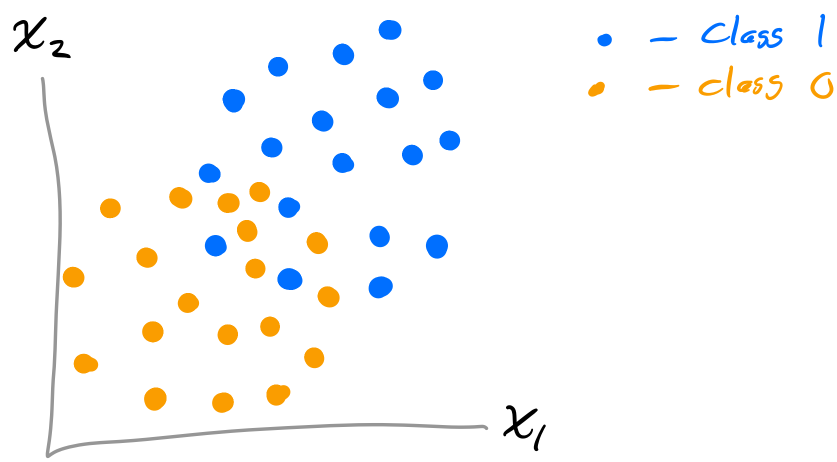
Training Data $\{\overset{\mathbb{R}^2}{x^{(i)}}, \overset{\mathbb{R}}{y_i}\}_{i=1\ldots n}$

$$y_i = \begin{cases} 1 & \text{if class } 1 \\ 0 & \text{if class } 0 \end{cases}$$



Given this data, draw a decision boundary (curve where you would say class 1 is on one side and class 2 is on the other side)



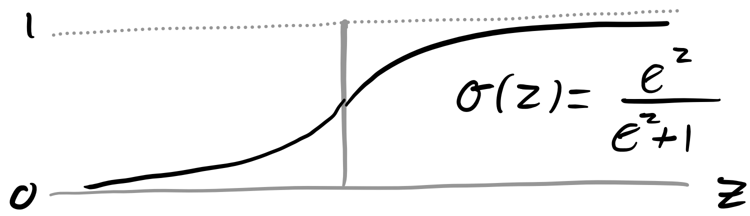


Model $$y = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2) = \hat{y}(x;\theta)$$

where

"sigmoid"

"logistic function"



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

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

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

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

Decision Boundary for Logistic Regression

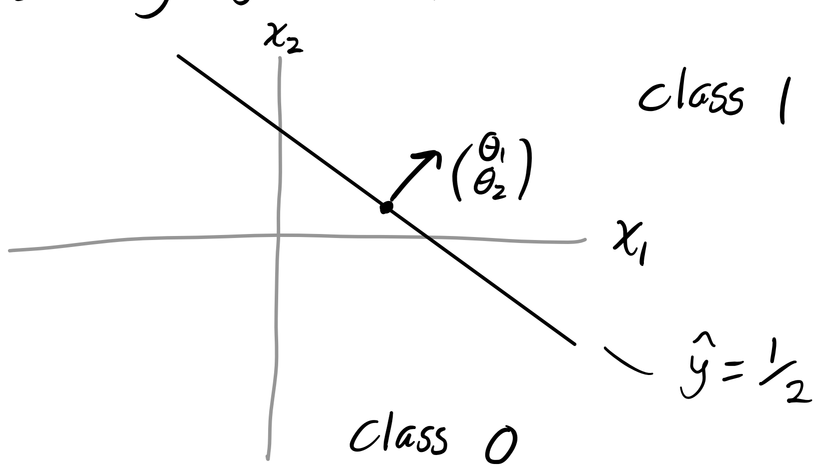
Training Data: $\left\{ \overset{\mathbb{R}^2}{x^{(i)}}, \overset{\mathbb{R}}{y_i} \right\}_{i=1 \dots n}$ $\qquad y_i = \begin{cases} 1 & \text{if class } 1 \\ 0 & \text{if class } 0 \end{cases}$

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

Predict: For new sample x, predict

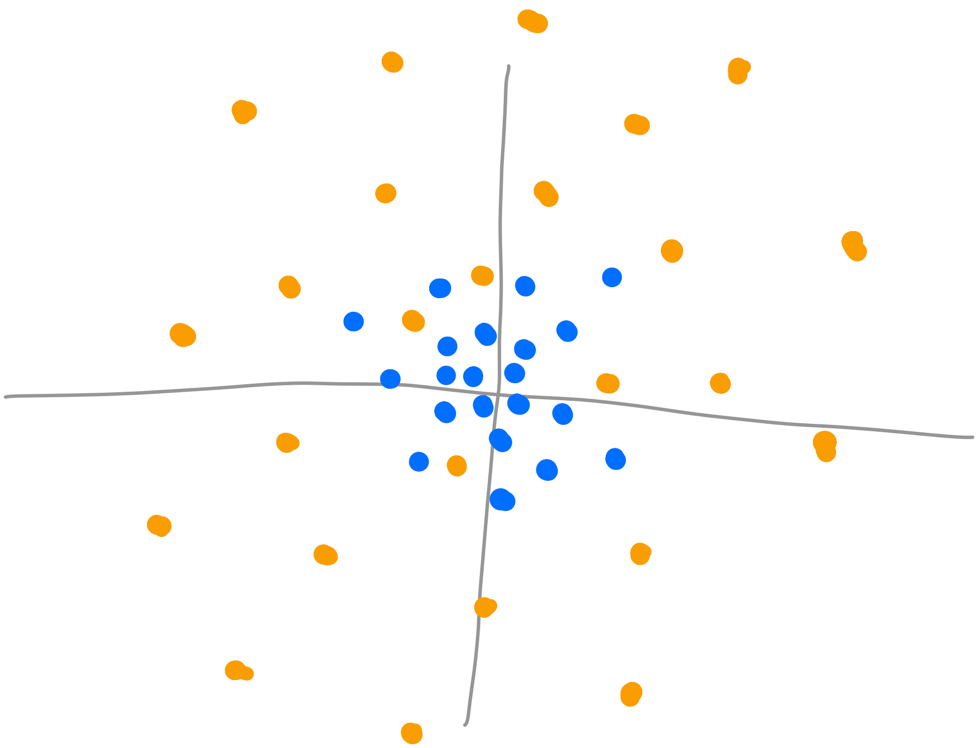
$$\begin{cases} \text{class } 1 & \text{if } \hat{y} \geq 1/2 \\ \text{class } 0 & \text{if } \hat{y} < 1/2 \end{cases}$$

Decision boundary is linear



Activity:

Could you use logistic regression to build a reasonable classifier for the following data?



Evaluating Classifiers

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

Accuracy: $\frac{TP + TN}{total}$

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

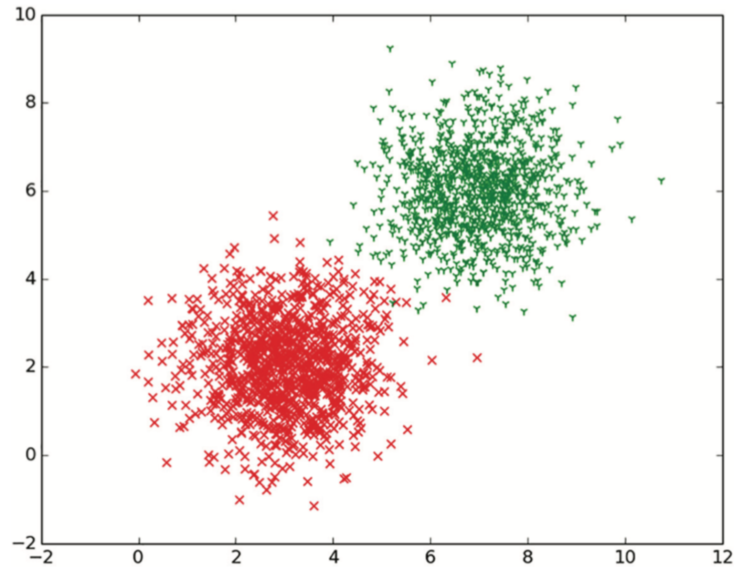
Training Data: $\{\underset{\mathbb{R}^2}{x^{(i)}}, \underset{\mathbb{R}}{y_i}\}_{i=1\ldots n}$ $\quad y_i = \begin{cases} 1 & \text{if class } 1 \\ 0 & \text{if class } 0 \end{cases}$

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

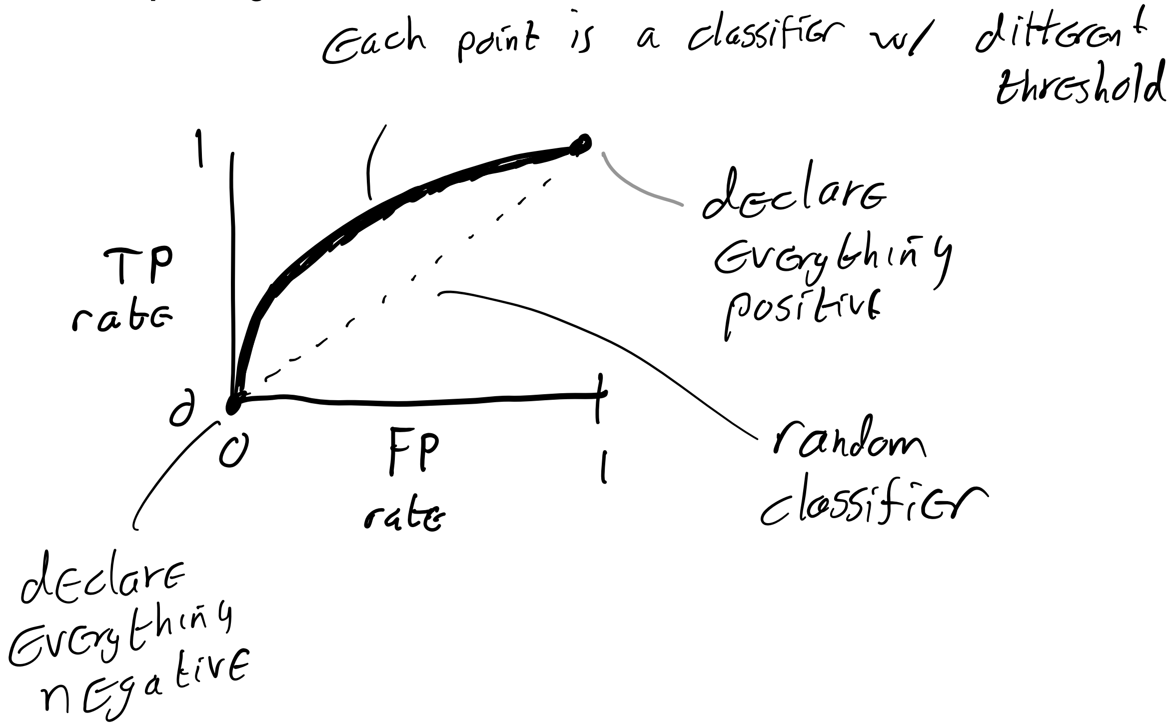
Predict: For new sample x, predict

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

could choose any value



Receiver Operating Characteristic Curves



Comparing classifiers and Area-Under-Curve (AUC)

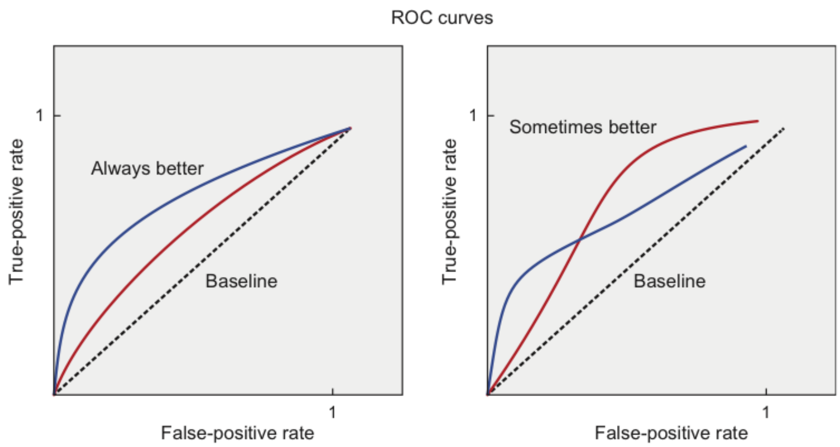
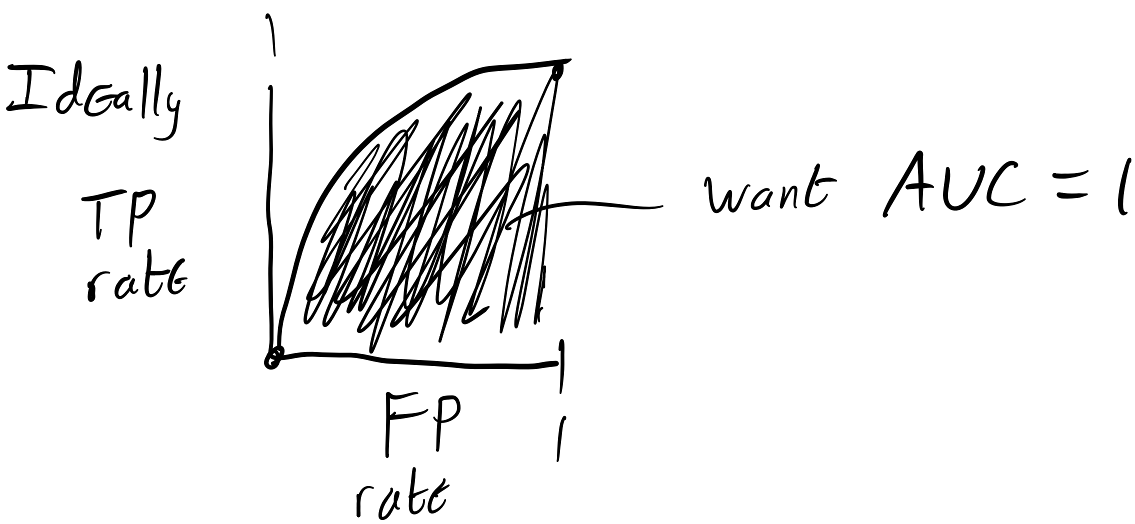


Figure 5.6 The principled way to compare algorithms is to examine their ROC curves. When the true-positive rate is greater than the false-positive rate in every situation, it's straightforward to declare that one algorithm is dominant in terms of its performance. If the true-positive rate is less than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



Also common to plot precision-recall curves