Logistic regression
DS351

Classification

Given features, want to predict **binary** or **categorical** variables

$$\{0, 1\}, \{\text{True}, \text{False}\}$$

$$\text{features} \Rightarrow \text{label}$$

$$\{1, 2, \ldots, N\}$$

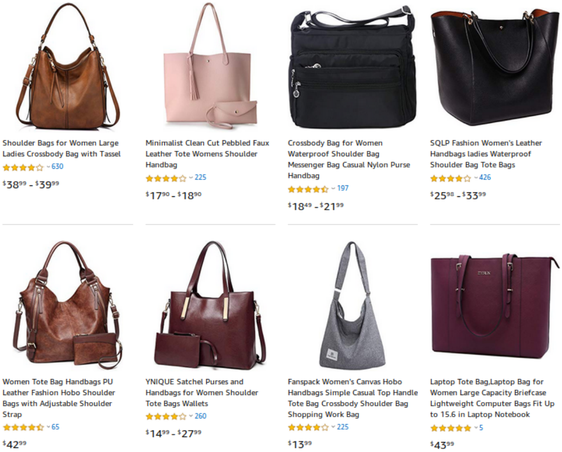
Classification problems

Is this a **cat** or a **dog**?

(e.g. robot, automatic car)

(cat)

Classification problems



Will I **click on** these products?

(no)

Probabilistic framework

Instead of directly predicting 0's and 1's

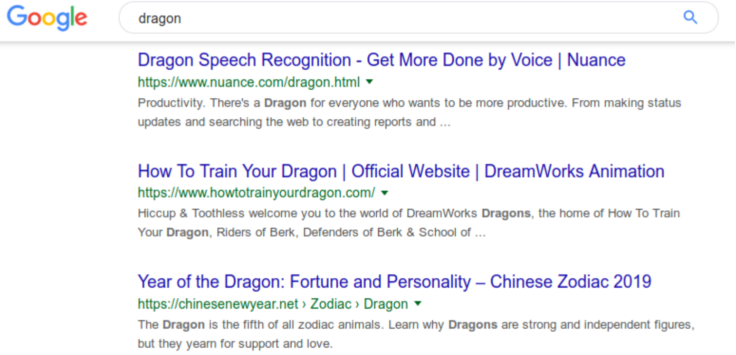
1	leopard
0	jaguar
0	cheetah
0	snow leopard
0	egyptian cat

we could predict the probability of being in each class:

0.724	leopard
0.181	jaguar
0.062	cheetah
0.03	snow leopard
0.003	egyptian cat

Applications

- **Ranking** of the search results by probabilities



- **Medical diagnosis**
  - Looking at the heart rate, blood pressure etc., what is the chance of contracting a heart disease?

Binary classification

Given: an instance with features $\boldsymbol{x}$ and possible label $y = 0$ or $y = 1$.
Goal: Predict the probability of the instance being in class 0 and 1:

$$P(y = 0|\boldsymbol{x}) \qquad \text{and} \qquad P(y = 1|\boldsymbol{x})$$

We then make the following prediction:

$$\boxed{\hat{y} = \begin{cases} \mathbf{0} & \text{if } P(y = 1|\boldsymbol{x}) \leq 0.5 \\ \mathbf{1} & \text{if } P(y = 1|\boldsymbol{x}) > 0.5 \end{cases}}$$

Multiclass classification

Given: an instance with features $\boldsymbol{x}$ and possible label $y = 1, 2, \ldots, N$.
Goal: Predict the probability of the instance being in class $1, 2, \ldots, N$:

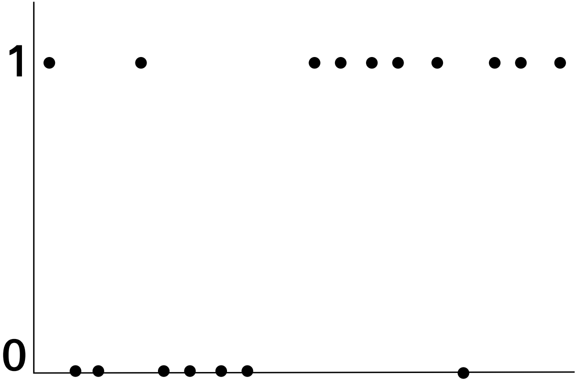
$$P(y = j|x) \quad \text{for } j = 1, 2, \ldots, N$$

We then make the following prediction:

$$\boxed{\hat{y} = J \text{ if } P(y = J|\boldsymbol{x}) > P(y = j|\boldsymbol{x}) \text{ for any other } j}$$

Predicting probability

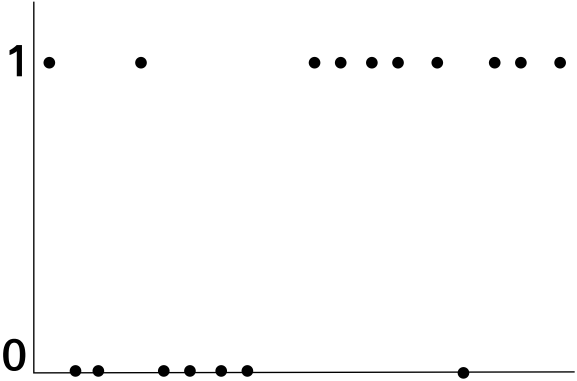
Can we use linear regression to do this?



We need some function that stays between 0 and 1.

Predicting probability

Instead, we need something like this:



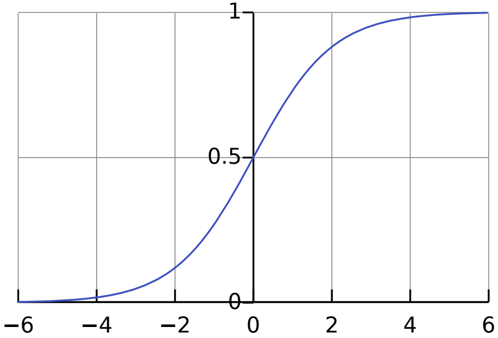
Logistic regression

That is, we are looking for a function with the following properties:

1. Stays between 0 and 1
2. Continuous
3. Symmetric

Logistic regression

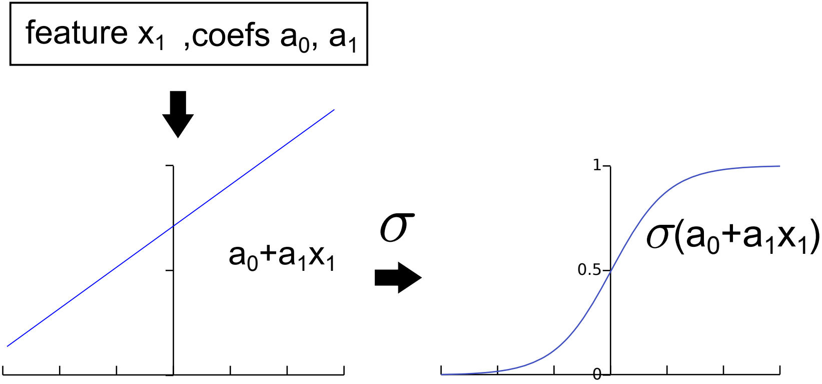
Sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$



- If $x \to -\infty$ then $\sigma(x) \to 0$.
- If $x \to \infty$ then $\sigma(x) \to 1$.

From features to probability

If the data is (x_1, y)

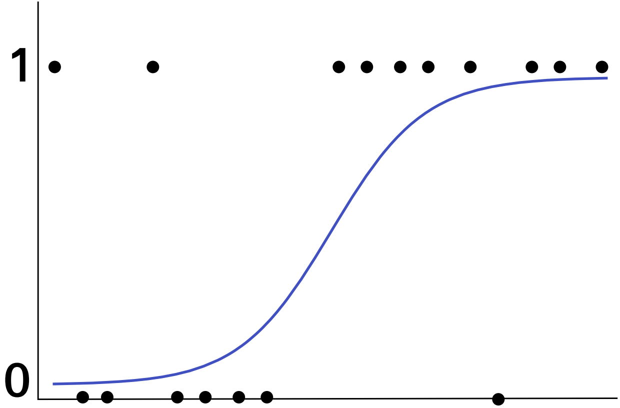


Logistic regression

Find *coefficients* $A = [a_0, a_1, \ldots, a_m]$ such that

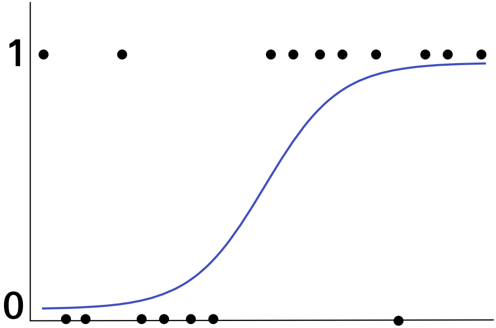
$$P(y = 1|x) = \frac{1}{1 + e^{-(a_0 + a_1 x_1 + \ldots + a_m x_m)}} = \frac{1}{1 + e^{-A \cdot \boldsymbol{x}}}$$

best fit the data



Logistic regression

$$P(y=1|x) = \frac{1}{1+e^{-(a_0+a_1x_1+\ldots+a_mx_m)}} = \frac{1}{1+e^{-A\cdot\boldsymbol{x}}}$$



- If $a_0 + a_1x_1 + \ldots + a_mx_m \to \infty$ then $\sigma(x) \to 1$.
- If $a_0 + a_1x_1 + \ldots + a_mx_m \to -\infty$ then $\sigma(x) \to 0$.

Logistic regression

Find *coefficients* $A = [a_0, a_1, \ldots, a_m]$ such that

$$P(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-(a_0 + a_1 x_1 + \ldots + a_m x_m)}} = \frac{1}{1 + e^{-A \cdot \boldsymbol{x}}}$$

best fit the data

What is $P(y = 0|\boldsymbol{x})$?

Log-odds

How can we interpret the linear function $a_0 + a_1 x_1 + \ldots + a_m x_m$ in this model?

$$\log \left(\frac{P(\boldsymbol{x})}{1 - P(\boldsymbol{x})} \right) =$$

Log-odds

How can we interpret the linear function $a_0 + a_1 x_1 + \ldots + a_m x_m$ in this model?

$$\log\left(\frac{P(\boldsymbol{x})}{1 - P(\boldsymbol{x})}\right) =$$

- This is called **log-odds** or **logit**.
- 1 unit increase in $\boldsymbol{x}$ = 1 unit increase in log-odds

Maximum-likelihood principle

Principle: If the data point (x, y) already appears in the data, then the probability $P(y|x)$ is high.

= [x_1,x_2,...,x_{784}]

Goal: **Maximize** the probability $P(y|x)$ **for all** data points (x, y).

Maximum-likelihood principle

Given data:

$$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)}), \quad y = 0 \text{ or } 1$$

Find $A = [a_0, a_1, a_2, \dots, a_m]$ that maximizes

$$\begin{aligned}
\text{Likelihood} &= \text{Probability that the data appear} \\
L(A) &= P(y^{(1)}|x^{(1)} \text{ and } y^{(2)}|x^{(3)} \text{ and} \dots \text{ and } y^{(n)}|x^{(n)}) \\
&= P(y^{(1)}|x^{(1)}) P(y^{(2)}|x^{(3)}) \dots P(y^{(n)}|x^{(n)})
\end{aligned}$$

Maximum-likelihood principle

Example: $A = [a_0, a_1, \ldots, a_{784}]$

Example: Credit card data

Is the user going to default on their credit card?
$y = 1$: default, $y = 0$: not default

	Coefficient	Std. error	Z-statistic	P-value
Intercept	10.8690	0.4923	22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	0.6468	0.2362	2.74	0.0062

Example: Credit card data

Is the user going to default on their credit card?
$y = 1$: default, $y = 0$: not default

	Coefficient	Std. error	Z-statistic	P-value
Intercept	10.8690	0.4923	22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	0.6468	0.2362	2.74	0.0062

- 1 baht increase in balance = 0.0055 unit increase in log-odds
- $Z = \frac{\hat{\beta}_i}{\mathrm{SE}(\hat{\beta}_i)}$.
- $H_0 : \beta_1 = 0$ is rejected; there is an association between `balance` and the probability of default

Predictions

	Coefficient	Std. error	Z-statistic	P-value
Intercept	10.8690	0.4923	22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	0.6468	0.2362	2.74	0.0062

$$\hat{p}(y = 1 | x_1 = 1,500, x_2 = 40, x_3 = 1)$$
$$= \frac{1}{1 + e^{-(-10.869 + 0.00574 \times 1{,}500 + 0.003 \times 40 - 0.6468 \times 1)}} = 0.058$$
$$\hat{p}(y = 1 | x_1 = 1,500, x_2 = 40, x_3 = 0)$$
$$= \frac{1}{1 + e^{-(-10.869 + 0.00574 \times 1{,}500 + 0.003 \times 40 - 0.6468 \times 0)}} = 0.105.$$

Non-students have higher chance of defaulting their cards.

Framingham dataset

- Label: Diagnosed with a heart disease in the next 10 years
- Features: gender, smoking, blood pressure, heart rate, blood sugar, cholesterol, BMI

The model

$$P(y = 1|\text{CigsPerDay, Chol, BMI ...})$$
$$= \frac{1}{1 + e^{-(0.04\text{CigsPerDay}+0.002\text{Chol}+0.003\text{BMI}+...)}}.$$

- If $P(y = 1|\text{CigsPerDay, Chol, BMI ...}) = 0.2 \Rightarrow$, classify y as 0
- If $P(y = 1|\text{CigsPerDay, Chol, BMI ...}) = 0.8 \Rightarrow$ classify y as 1
- With everything else fixed, higher CigsPerDay $\Rightarrow$ higher chance of heart disease.
- $+1$ cigarette per day $= +0.04$ log-odds.

Cross-validation accuracy

$$\text{Accuracy} = \frac{\#\text{Correctly classified}}{\#\text{Total}}$$

10-fold cross-validation

- Split data into 10 sets.
- Train on 9 sets, test on the remaining set.
- Keep changing the test set. Obtain 10 accuracies.
- Compute the average accuracy.

	1NN	3NN	5NN	7NN	9NN	Logistic
Accuracy	77.55	81.96	83.18	83.96	84.29	85.40

Multiclass logistic regression

Similar to the binary logistic regression: given features x, the probability of a data point being in each group is defined by

$$P(y=1|\boldsymbol{x}) = \frac{e^{A_1\cdot\boldsymbol{x}}}{1+\sum_{i=1}^{n-1} e^{A_i\cdot\boldsymbol{x}}}$$

$$P(y=2|\boldsymbol{x}) = \frac{e^{A_2\cdot\boldsymbol{x}}}{1+\sum_{i=1}^{n-1} e^{A_i\cdot\boldsymbol{x}}}$$

$$\cdots$$

$$P(y=N-1|\boldsymbol{x}) = \frac{e^{A_{N-1}\cdot\boldsymbol{x}}}{1+\sum_{i=1}^{n-1} e^{A_i\cdot\boldsymbol{x}}}$$

$$P(y=N|\boldsymbol{x}) = \frac{1}{1+\sum_{i=1}^{n-1} e^{A_i\cdot\boldsymbol{x}}}$$

Now we want to find $A_1, A_2, \cdots, A_{N-1}$ that maximizes the likelihood.

Example: number of parameters

Suppose we want to classify 28×28 images into three groups.

Example

When we use the model after training:

$\boldsymbol{x}$:

- If

  $$P(y = 1|\boldsymbol{x}) = 0.3, P(y = 2|\boldsymbol{x}) = 0.3, P(y = 3|\boldsymbol{x}) = 0.4$$

  classify $y = 3$.
- If

  $$P(y = 1|\boldsymbol{x}) = 0.2, P(y = 2|\boldsymbol{x}) = 0.4, P(y = 3|\boldsymbol{x}) = 0.4$$

  randomly pick $y = 2$ or $y = 3$.