Lecture 5 – Supervised Learning

Classification models

Information Systems for Sustainable Society (is3) | WiSo Faculty | Univ.-Prof. Dr. Wolfgang Ketter | 11/11/22

Agenda

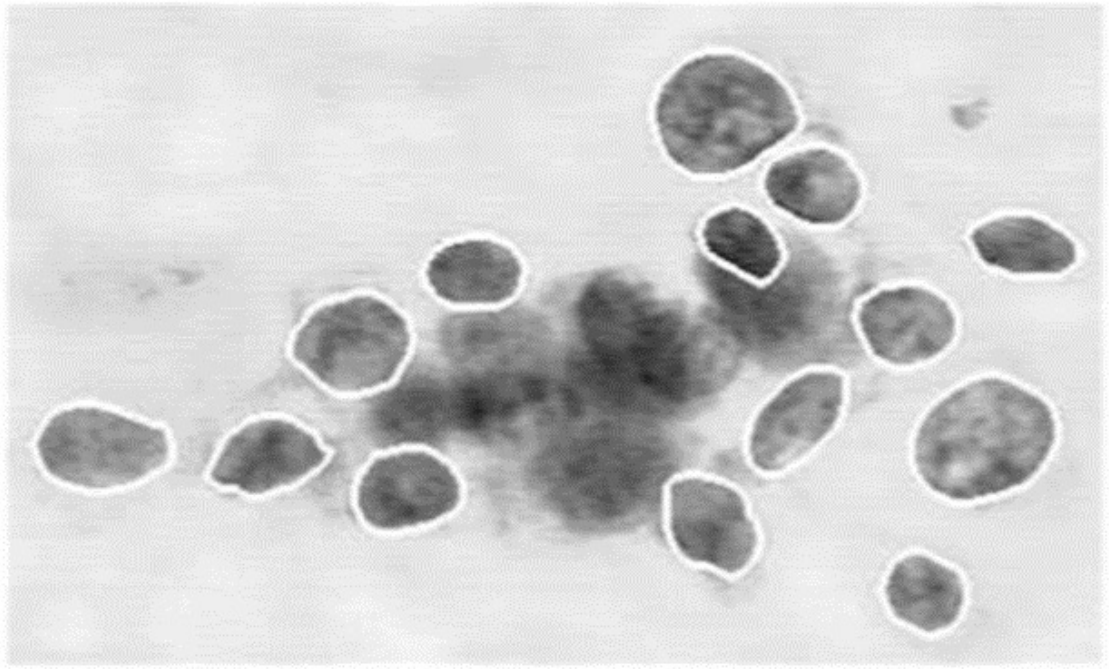
- Linear Classifiers (continue)
- Classification via Logistic Regression
- Naïve Bayes Classifier
- Review probability

To develop an understanding of classification we will work with a concrete example -**Breast cancer dataset**

Breast cancer dataset

- Setting: researchers took **569 images of cancerous cells**, under a microscope, and manually selected the outlines of the different cells (this step is the kind of thing that would ideally be replaced by automatic computer vision architectures in current systems)
- Researchers then considered **10 different features of each cell**, of instance the area, perimeter, texture, number of concave points (i.e., indentations), variance of grayscale color, and some others

Extract



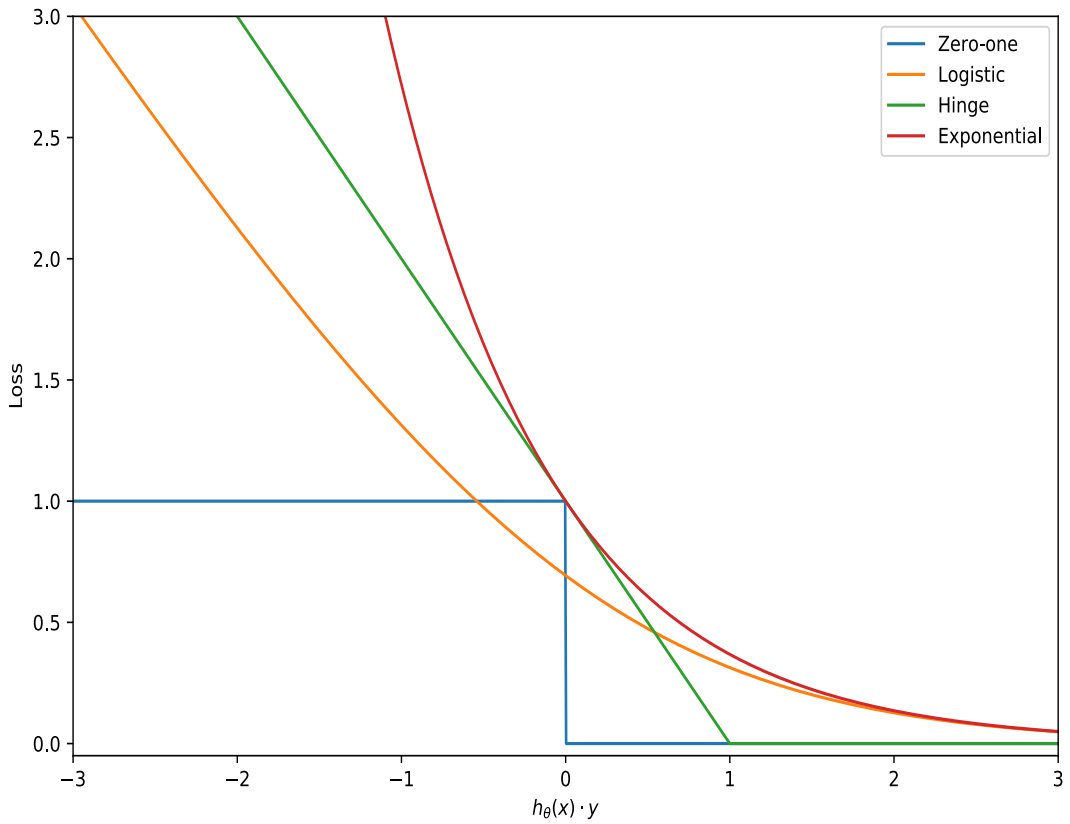
Let us again use our common ML terminology to formally define the classification setting

	Term	Instantiation in cancer domain
Input features	$x^{(i)} \in \mathbb{R}^n, i = 1, \dots, m$	$x^{(i)} = \begin{bmatrix} 1 \\ Mean - Area^{(i)} \\ Mean - Concave - Points^{(i)} \end{bmatrix}$
Outputs	$y^{(i)} \in \mathcal{Y}, i = 1, \dots, m$	$y^{(i)} \in \{-1\ (benign), +1\ (malignant)\}$
Model parameters	$\theta \in \mathbb{R}^n$	$\theta = (\theta_0, \theta_1, \theta_2)$
Hypothesis function	$h_\theta: \mathbb{R}^n \rightarrow \mathbb{R}$	$h_\theta(x) = \theta^T x; \quad \hat{y} = sign(h_\theta(x))$
Loss function	$\ell: \hat{\mathcal{Y}} \times \mathcal{Y} \rightarrow R_+$	$\ell = \exp(-y\, h_\theta(x))$

Solving classification tasks: Machine learning optimization

- With this notation, the “canonical” machine learning problem is written in the exact same way
  - $\underset{\theta}{Minimize} \sum_{i=1}^{m} \ell\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right)$
- Unlike least squares, there is not an analytical solution to the zero gradient condition for most classification losses
- Instead, we solve these optimization problems using gradient descent (or a alternative optimization method, but we’ll only consider gradient descent here)
  - $Repeat\text{: } \theta := \theta - \alpha \sum_{i=1}^{m} \nabla_\theta \ell\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right)$

Overview of linear loss functions



1. **0/1 loss**

$$\ell_{0/1}(h_\theta(x), y) = \begin{cases} 0 \; if sign(h_\theta(x)) = y \\ 1 \; otherwise \end{cases}$$
$$= 1\{y \; h_\theta(x) \leq 0\}$$

2. **Exponential loss**

$$\ell_{\text{exp}} = \exp(-\text{y} \; \text{h}_\theta(\text{x}))$$

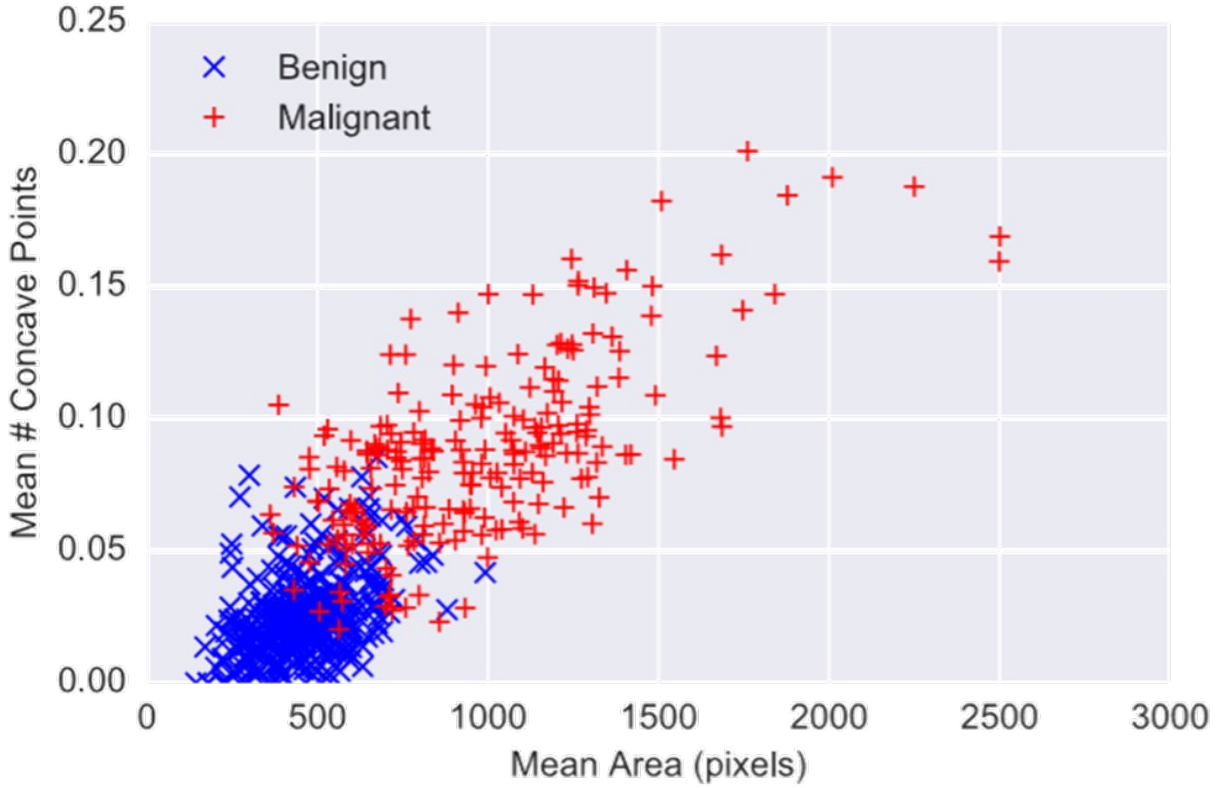
3. **Logistic loss**

$$\ell_{\text{logistic}} = \log\big(1 + \exp\big(-\text{y} \; \text{h}_\theta(\text{x})\big)\big)$$

4. **Hinge loss**

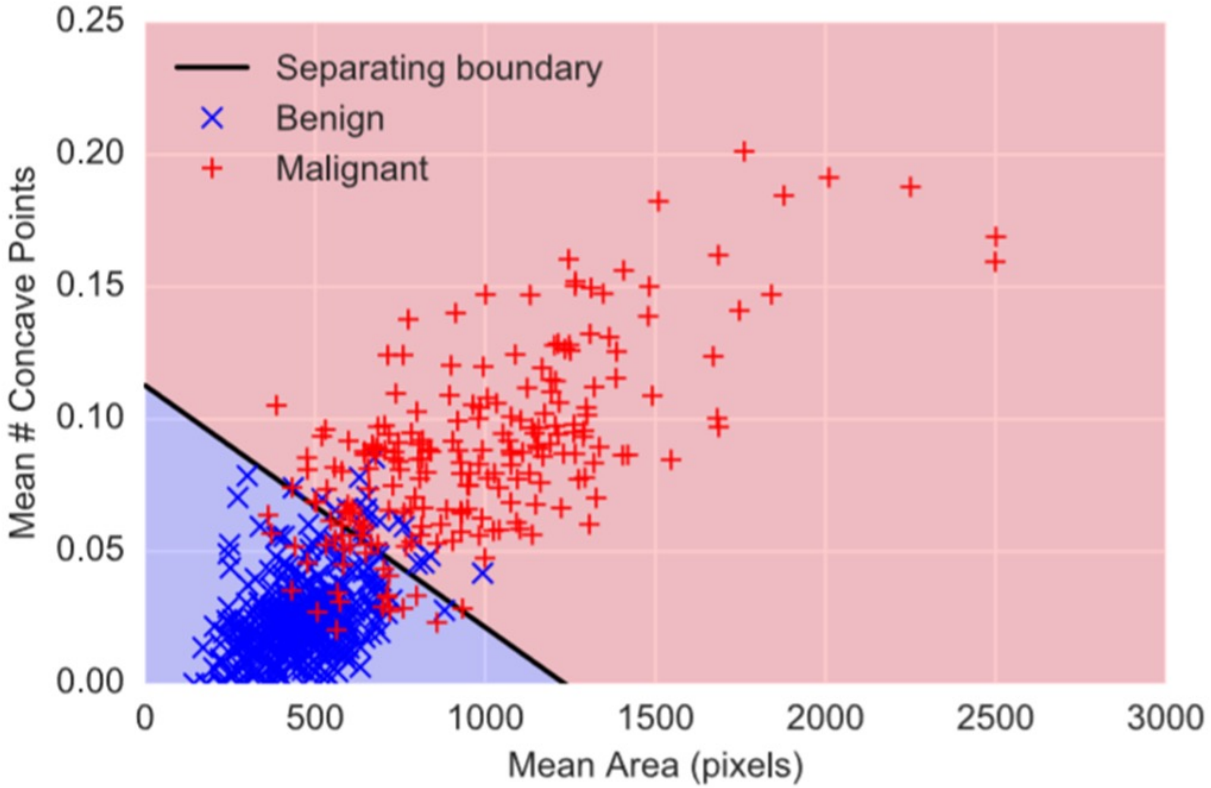
$$\ell_{\text{hinge}} = \max\{1 - \text{y} \; \text{h}_\theta(\text{x}), 0\}$$

Example: breast cancer classification



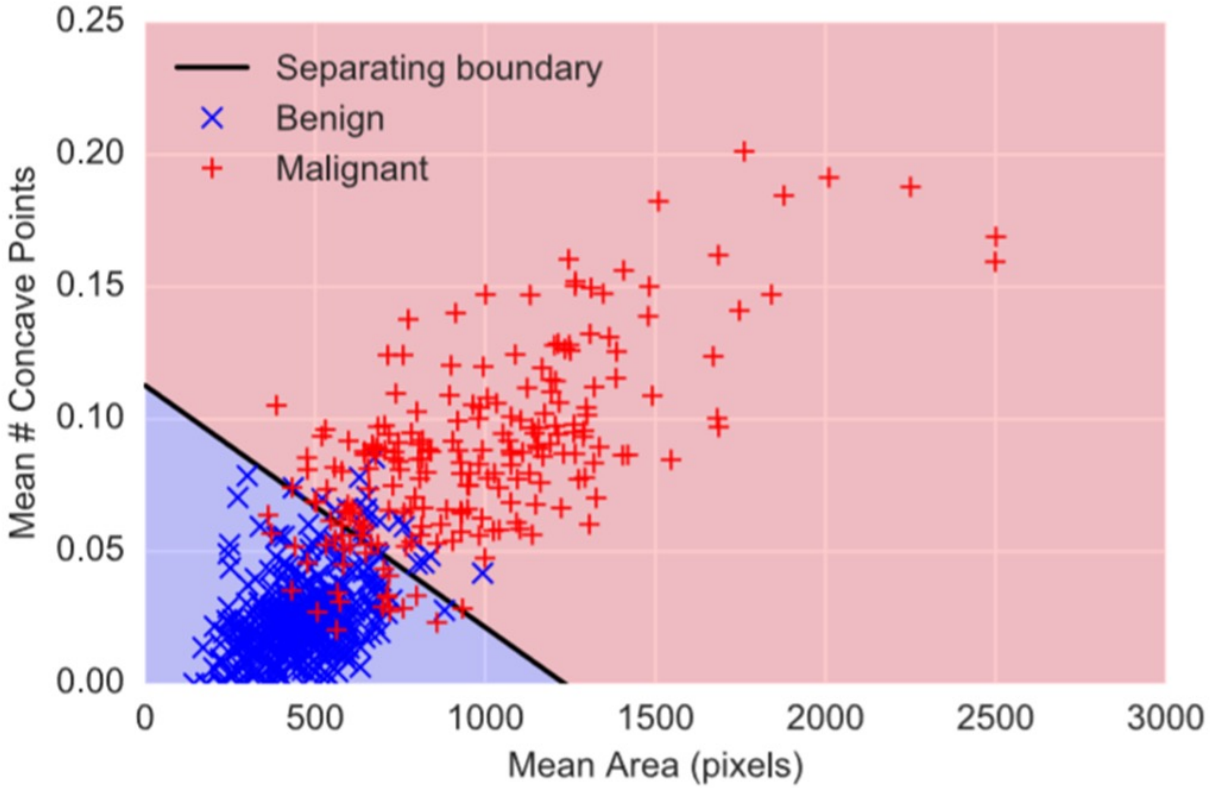
- Plot of two features: mean area vs. mean concave points, for two classes
- There is obviously some structure to the data: cells with greater average area and greater numbers of concave points are more likely to be malignant

Linear classification example



- Linear regression ≡ “fitting a line to the data”
- **Linear classification ≡ “separating the classes with a line”**

Understanding linear classification diagrams

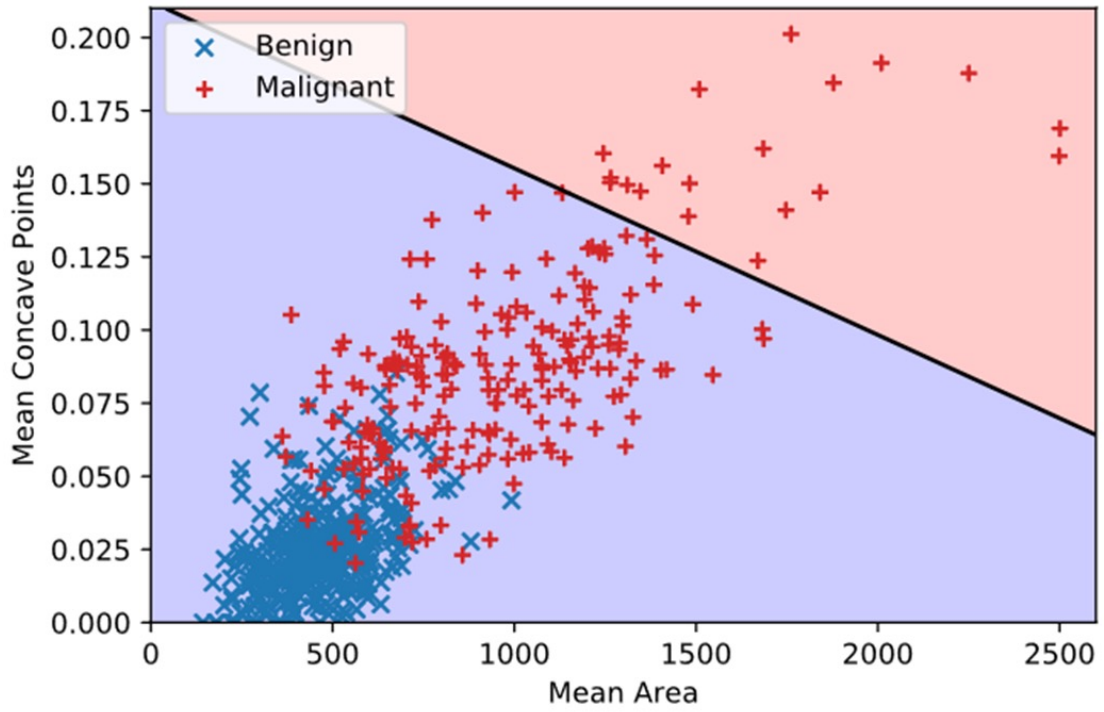


- Color shows region where the $h_\theta(x)$ is positive
- Separating boundary is given by the equation $h_\theta(x) = 0$

Support vector machine (SVM) – Linear hypothesis function with hinge loss

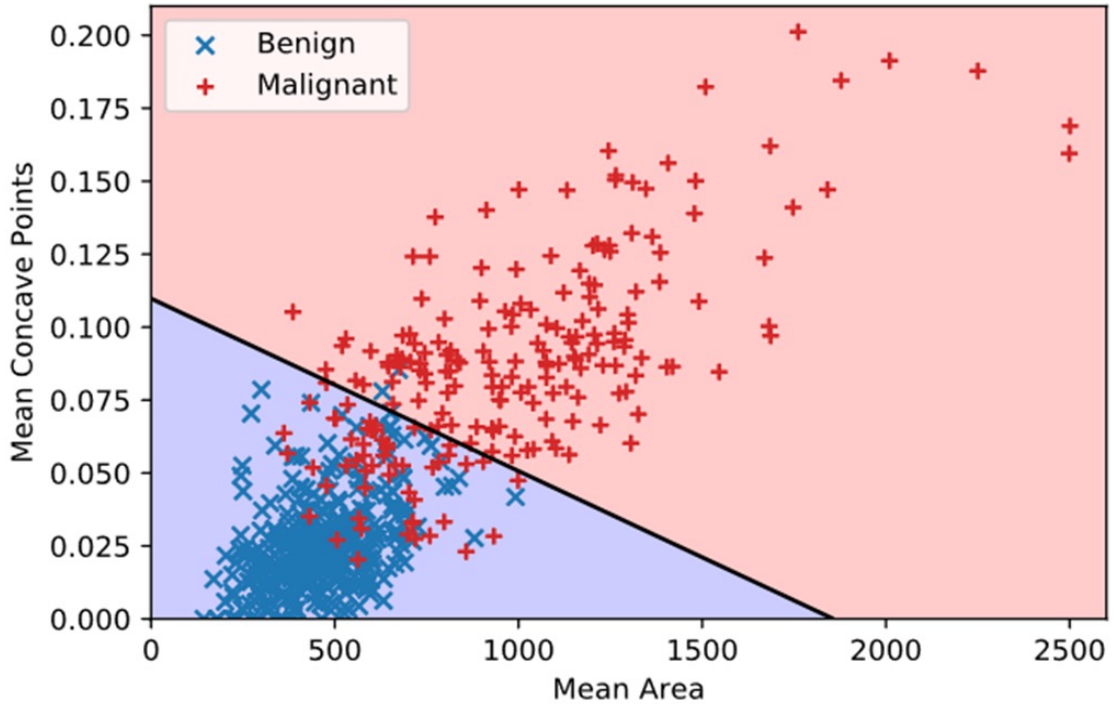
- A (linear) support vector machine (SVM) just solves the canonical machine learning optimization problem using **hinge loss** and **linear hypothesis**, **plus** an additional **regularization** term
  - $\underset{\theta}{Minimize} \sum_{i=1}^{m} max\left\{1 - y^{(i)}\theta^T x^{(i)}, 0\right\} + \frac{\lambda}{2} \|\theta\|_2^2$
- Even more precisely, the “standard” SVM does not actually regularize the "θ_0" (corresponding to the constant feature, but we’ll ignore this here)
- Updates using gradient descent:
  - $\theta := \theta - \alpha \sum_{i=1}^{m} -y^{(i)} x^{(i)} 1\left\{y^{(i)}\theta^T x^{(i)} \leq 1\right\} - \alpha\lambda\theta$

Support vector machine example



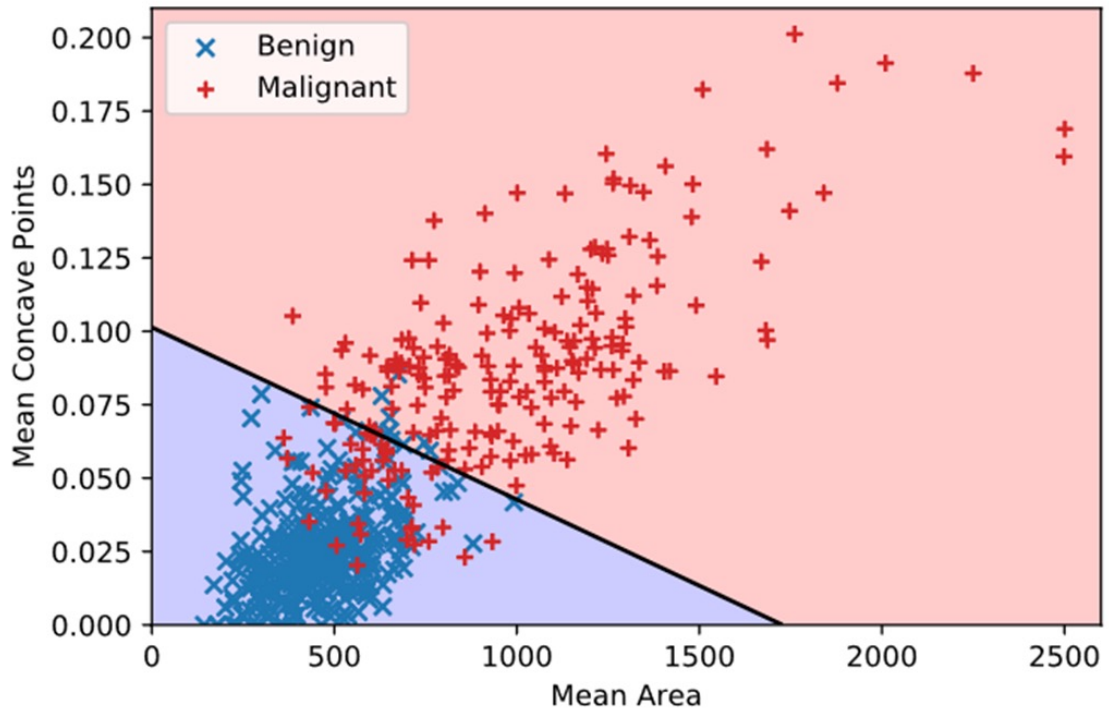
- Iterations: 10

Support vector machine example



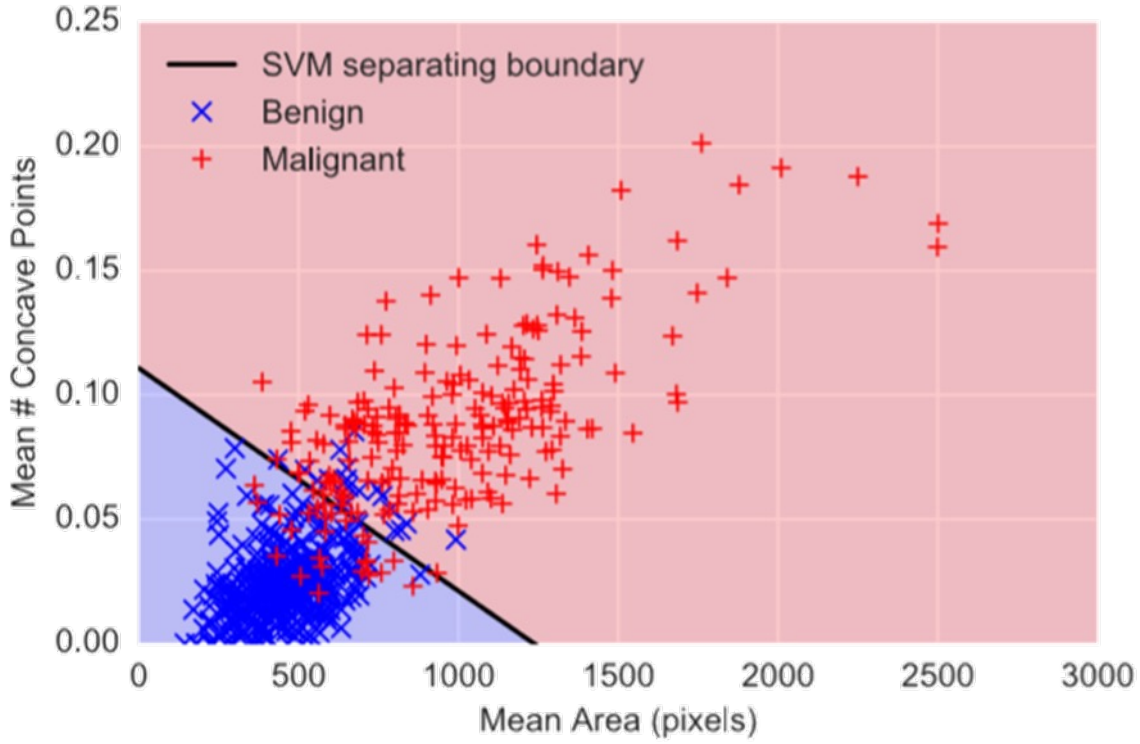
- Iterations: 50

Support vector machine example



- Iterations: 100

Support vector machine example

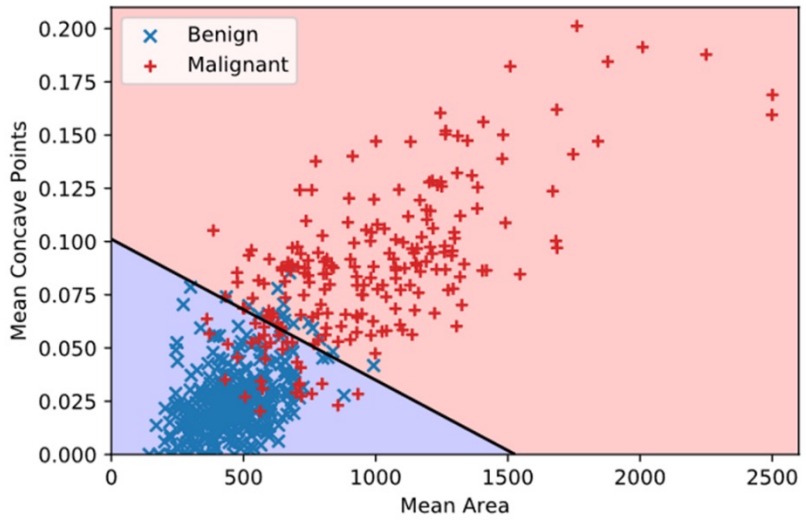


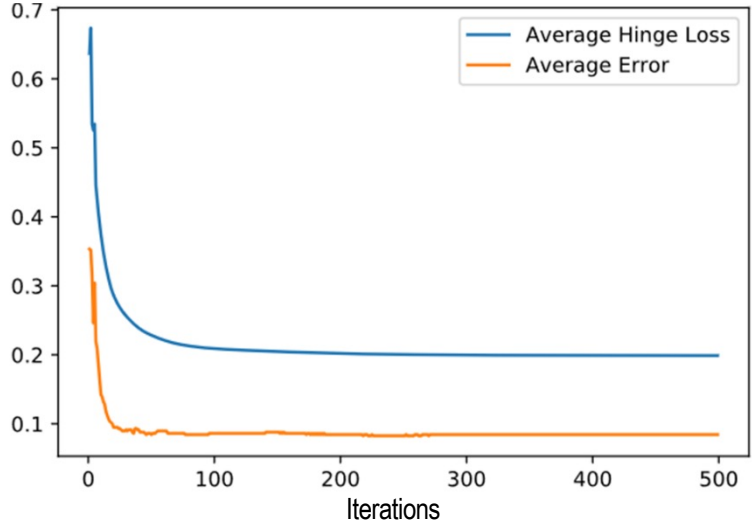
- Running support vector machine on cancer dataset, with small regularization parameter (effectively zero)
  - $\theta = \begin{bmatrix} 1.456 \\ 1.848 \\ -0.189 \end{bmatrix}$

How to pre-process the data for classification? – A look at some alternative normalization techniques

- Thus far, we have been normalizing all columns of the data (except for the constant feature), to lie in the range [0,1]. Just to highlight two common alternatives:
  1. **Normalize features to lie in the range [−1,1]**
  2. **Normalize them to have zero mean and unit variance**
- Let's look at the resulting behavior of the classifier for these two alternatives

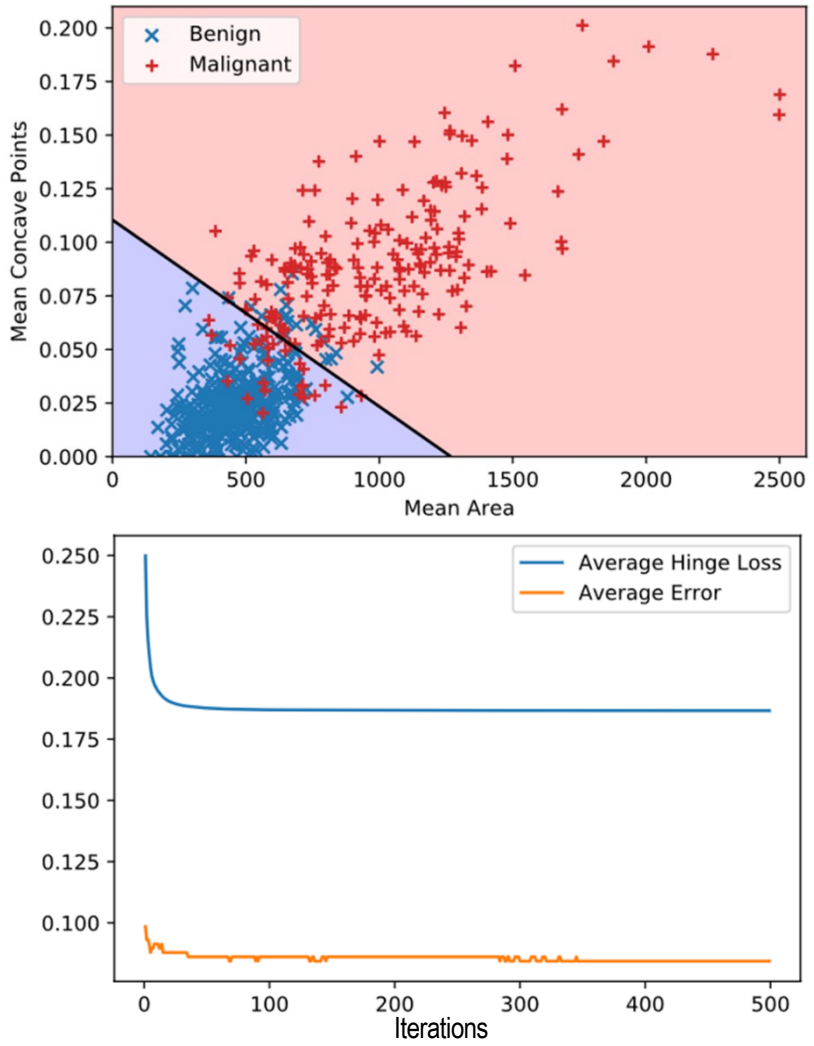
[−1,1] feature re-scaling





- [-1,1] is another common form of re-scaling method
- Let us look at what this means for the performance of SVM classification (hinge loss and classification error)
- Note: classification error is the ratio of wrongly classified samples vs. the full sample size

Zero mean and unit variance re-scaling



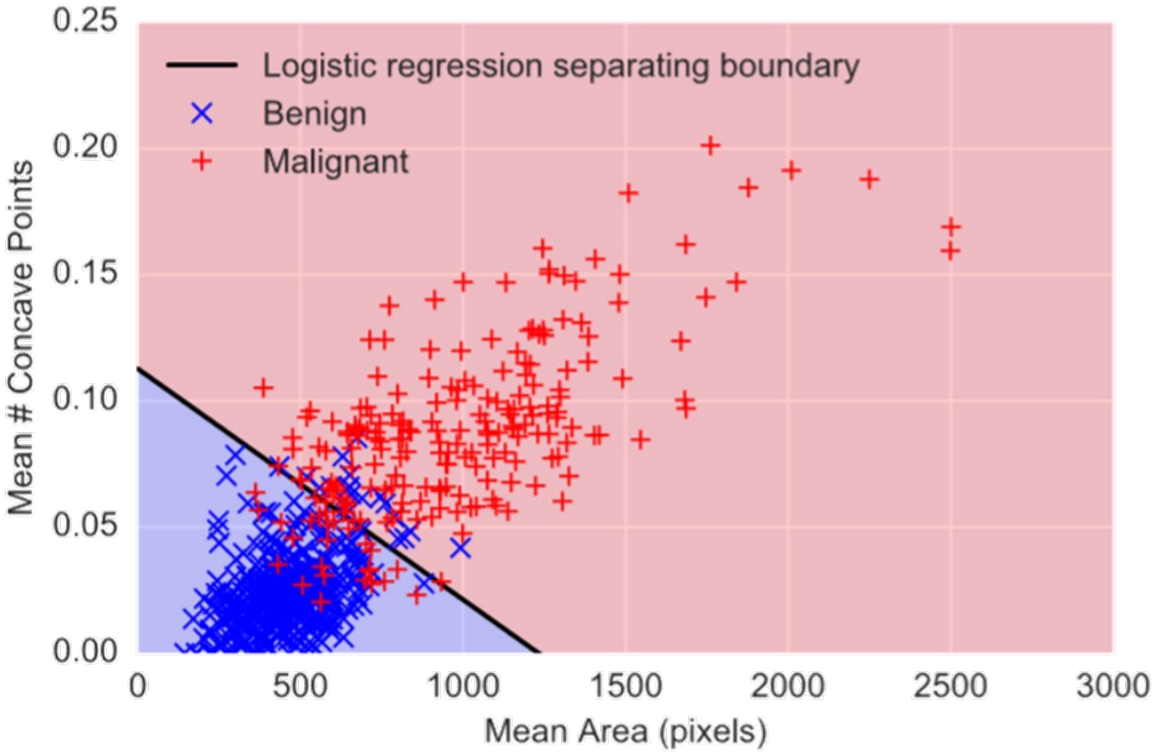
- Normalizing features to have zero mean is another common re-scaling technique for classification
- Indeed, it looks much better in this case – After just one gradient descent step, our classifier already has less than 10% error
- While the best re-scaling method is ultimately data-dependent, this strategy (normalizing columns to have zero mean and unit variance), is the most common strategy used in practice, and should be the default strategy you attempt when needing to normalize features for classification

Logistic regression – Linear hypothesis function and logistic loss

- Logistic regression just solves this problem using logistic loss and linear hypothesis function
  - $\underset{\theta}{Minimize} \sum_{i=1}^{m} \log(1 + \exp(-y^{(i)}\theta^T x^{(i)}))$
- Gradient descent updates (can you derive these?):
  - $\theta := \theta - \alpha \sum_{i=1}^{m} -y^{(i)}x^{(i)} \frac{1}{1+\exp(y^{(i)}\theta^T x^{(i)})}$
- Logistic regression also has a nice probabilistic interpretation: certain quantities give the *probability*, under a particular model, of an example being positive or negative

We will consider this probabilistic setting shortly!

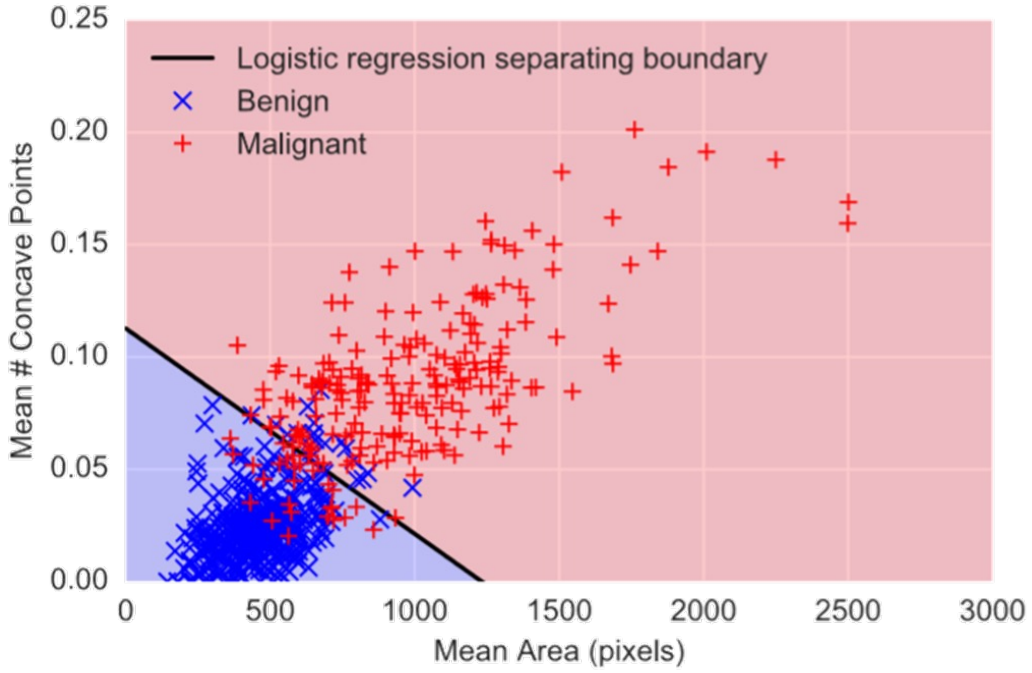
Logistic regression example



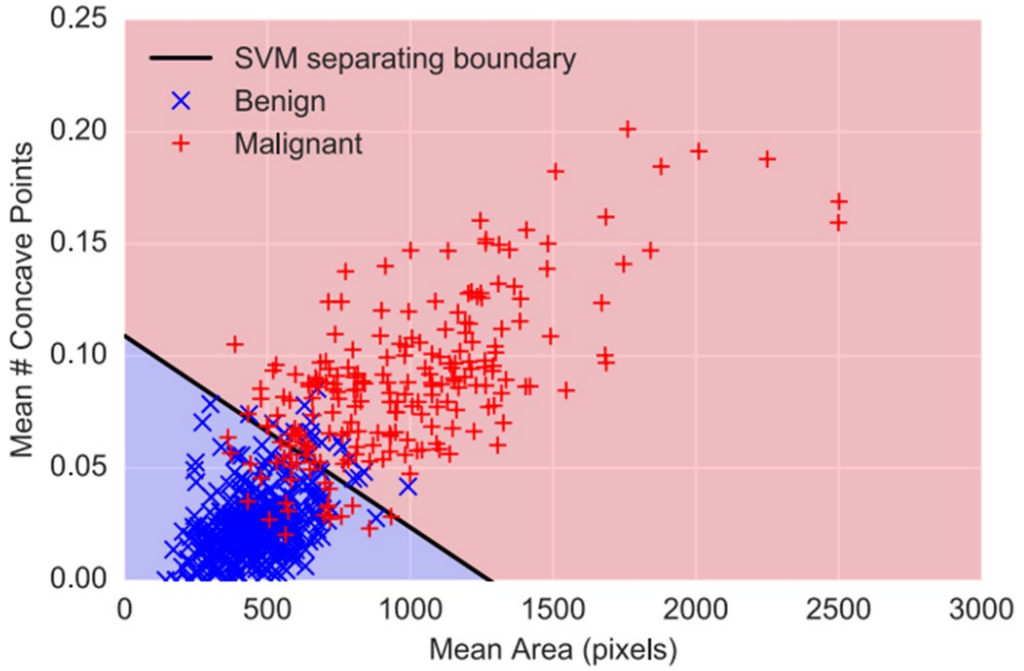
- Running logistic regression on cancer data set (small regularization)

Logistic regression and SVM regression yield very similar results – Why?

Logistic Regression

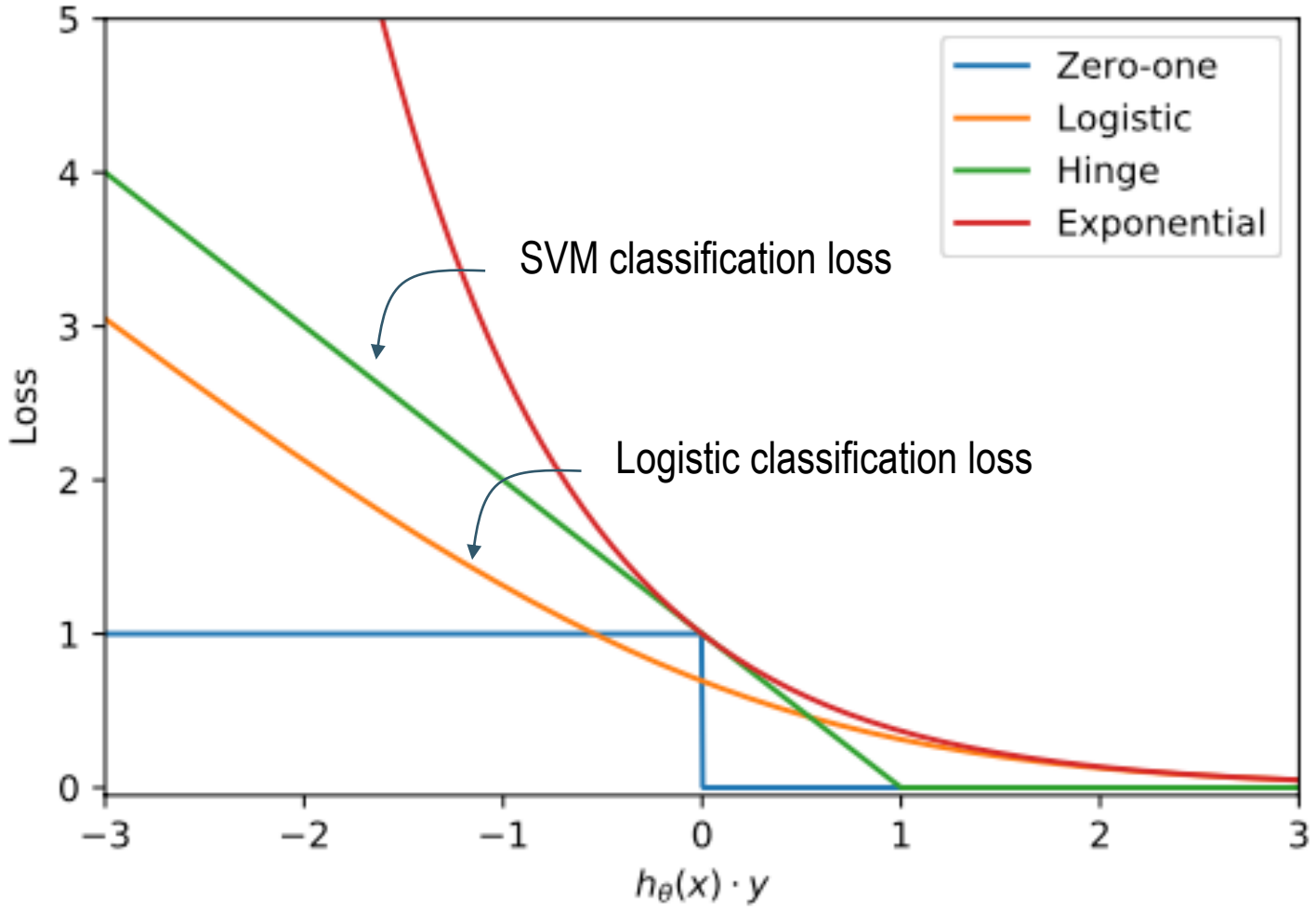


SVM



Why do logistic regression and SVM regression yield very similar results on the breast cancer dataset?

Logistic regression and SVM regression yield very similar results – This is because of the similar functional form of the



- SVM uses hinge loss
- Logistic regression uses logistic loss
- Both loss functions share similar properties: they both approach zero for $h_\theta(x) \cdot y$ large positive (hinge loss actually attains the zero value), and they both are approximately linear for $h_\theta(x) \cdot y$ large negative

Agenda

- Linear Classifiers (continue)
- **Classification via Logistic Regression**
- Review probability
- Naïve Bayes Classifier

Logistic Regression

- Logistic regression extends the ideas of linear regression to situation where the dependent variable Y is categorical. We can think of categorical variable as dividing the observation into classes.

- For example if Y denotes a recommendation on holding/ selling/ buying a stock, we have a categorical variable with three categories.

Logistic Regression is commonly used for classification and profiling tasks

- **Classification:**

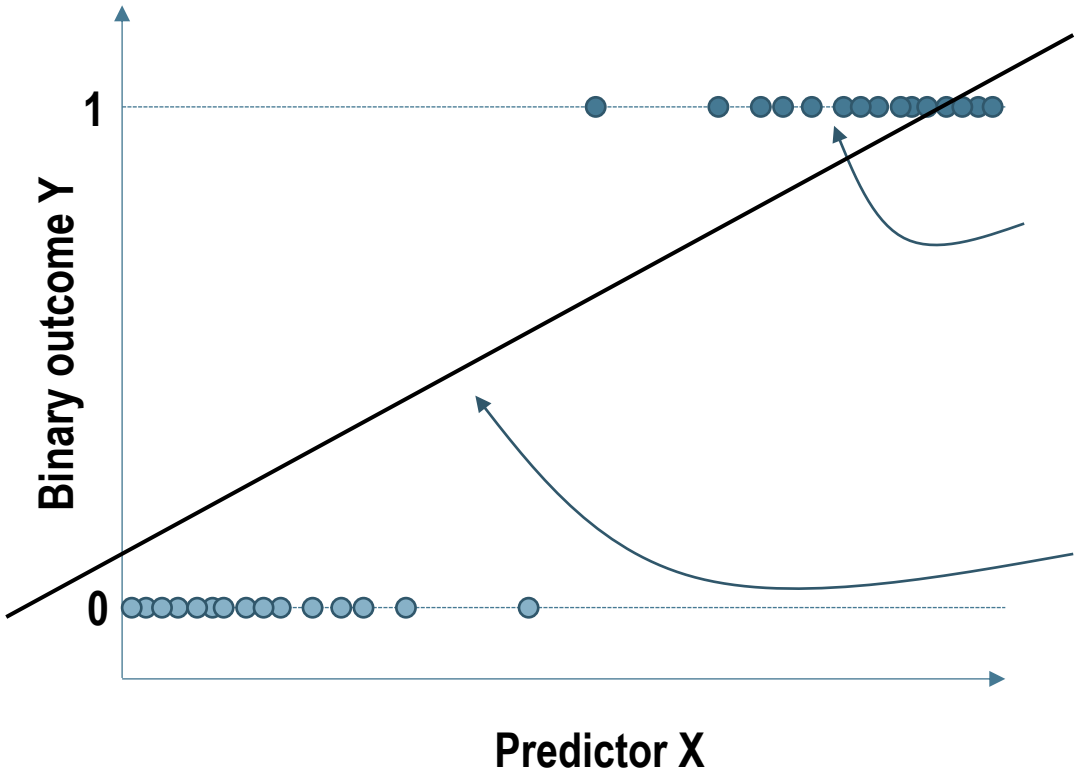
  Logistic regression can be used for classifying a new observation, where its class is unknown, into one of the classes, based on the values of its predictor

- **Profiling:**

  Logistic regression can also be used in data where the class is known, to find factors distinguishing between observations in different classes in terms of their predictor variables, or “predictor profile”. E.g., In the financial sector, institutions use profiling technologies for fraud prevention and credit scoring.

Logistic Regression – Example

Simple Linear Regression



Observations:
Dots indicate observations

Regression:
Standard linear regression

- **Problem:** extending the regression line a few units left or right along the X axis produces predicted probabilities that fall outside of [0,1]
- p = probability of belonging to class 1
- Need to relate p to predictors with a function that guarantees $0 \leq p \leq 1$
- Standard linear function does not

Which of the following statements is true?

a) If linear regression does not work on a classification task, applying feature scaling may help

b) If training set satisfies $0 \leq x^{(i)} \leq 1$ for every training example $(x^{(i)}, y^{(i)})$, then linear regression's prediction will also satisfy $0 \leq h_\theta(x) \leq 1$ for all values of x

c) None of the above.

Which of the following statements is true?

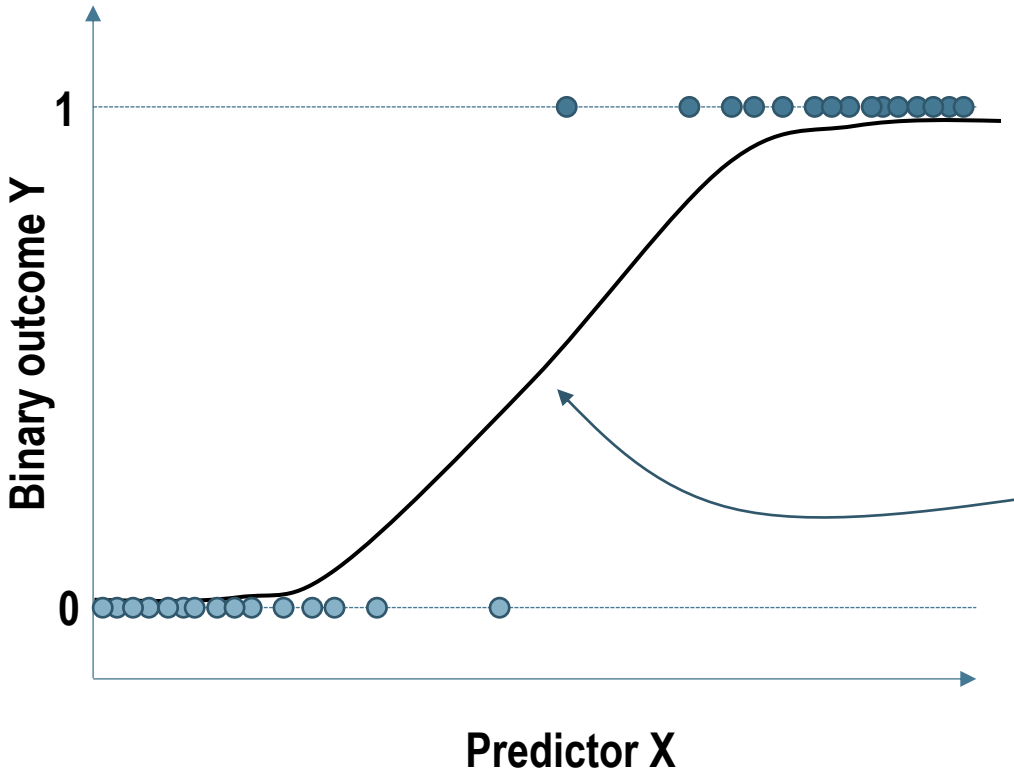
a) If linear regression does not work on a classification task, applying feature scaling may help

b) If training set satisfies $0 \leq x^{(i)} \leq 1$ for every training example $(x^{(i)}, y^{(i)})$, then linear regression's prediction will also satisfy $0 \leq h_\theta(x) \leq 1$ for all values of x

c) None of the above.

A Better Solution

Logistic Regression



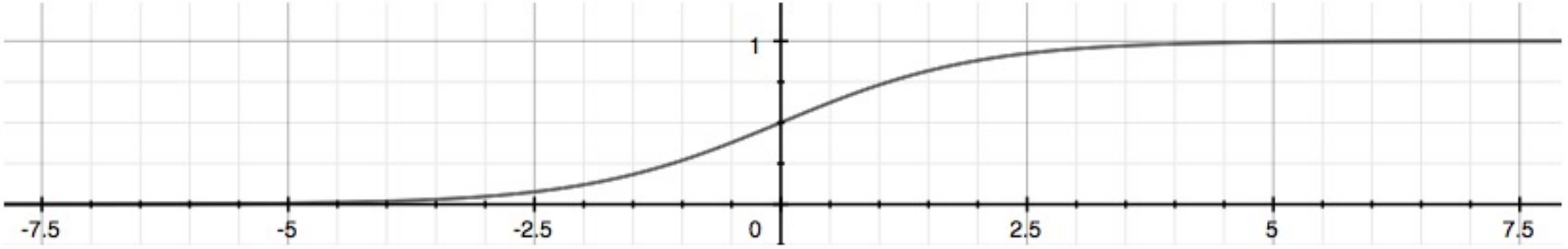
Regression:
Sigmoid function!
(bounded by asymptotes y=0 and y=1)

- **Solution:** use a sigmoidal regression function that is bounded by asymptotes y=1 and y=0

Logistic Regression Model

- We want $0 \leq h_\theta\ (x) \leq 1$:

$$h_\theta(x) = g(\theta^T x)$$
$$z = \theta^T x$$
$$g(z) = \frac{1}{1 + e^{-z}}$$



Background

- The idea behind logistic regression is straightforward: instead of using Y as the dependent variable, we use a function of it called the logit.
- Given some event with probability p of being 1, the odds of that event are given by:

$$Odds = \frac{p}{1-p}$$

- Consider the following data:

		Delinquent		Total
		Yes	No	
Testosterone	Normal	402	3614	4016
	High	101	345	446
Total		503	3959	4462

- The odds of being delinquent if you are in the Normal group are:

P(delinquent)/(1–p(delinquent)) =
(402/4016) / (1 - (402/4016)) =
0.1001 / 0.8889 = 0.111

Odds Ratio

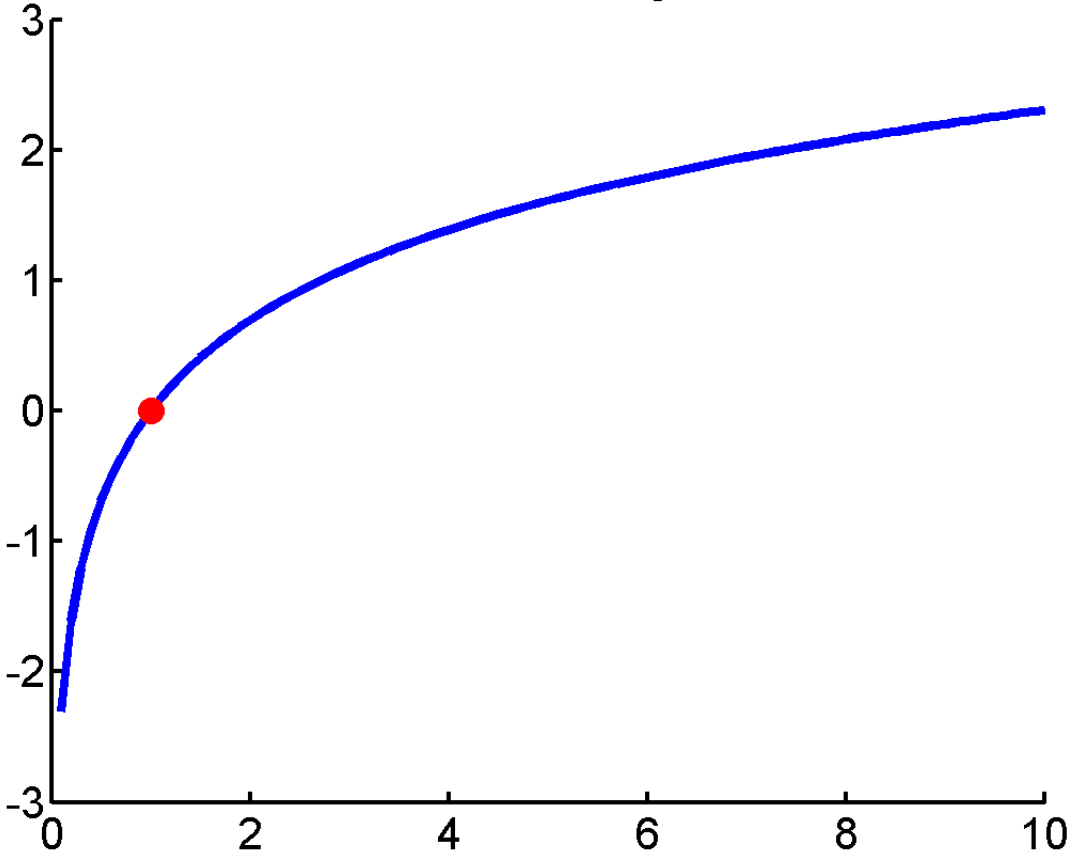
- The odds of being not delinquent in the Normal group is the reciprocal of this:
  - 0.8999/0.1001 = 8.99
- Now, for the High testosterone group
  - odds(delinquent) = 101/345 = 0.293
  - odds(not delinquent) = 345/101 = 3.416
- When we go from Normal to High, the odds of being delinquent nearly triple:
  - Odds ratio: 0.293/0.111 = 2.64
  - 2.64 times more likely to be delinquent with high testosterone levels

- Consider the following data:

		Delinquent		Total
		Yes	No	
Testosterone	Normal	402	3614	4016
	High	101	345	446
Total		503	3959	4462

Logit Transform

- Logit(p) = ln(odds) = ln $(\frac{p}{1-p})$



- The logit is the natural log of the odds

Logistic Regression

- In logistic regression, we seek a model:

$$logit(p) = \beta_0 + \beta_1 X$$

- That is, the log odds (logit) is assumed to be linearly related to the independent variable X
- So, now we can focus on solving an ordinary (linear) regression!

Logistic Regression Model

- In logistic regression we take two steps:

1. **Step 1**: The first step yields estimates of the probabilities of belonging to each class. In the binary case we get an estimate of p = P(Y=1), the probability of belonging to class 1 (which also tells us the probability of belonging to class 0)
2. **Step 2**: In the next step we use a cutoff value on these probabilities in order to classify each case into one of the classes.

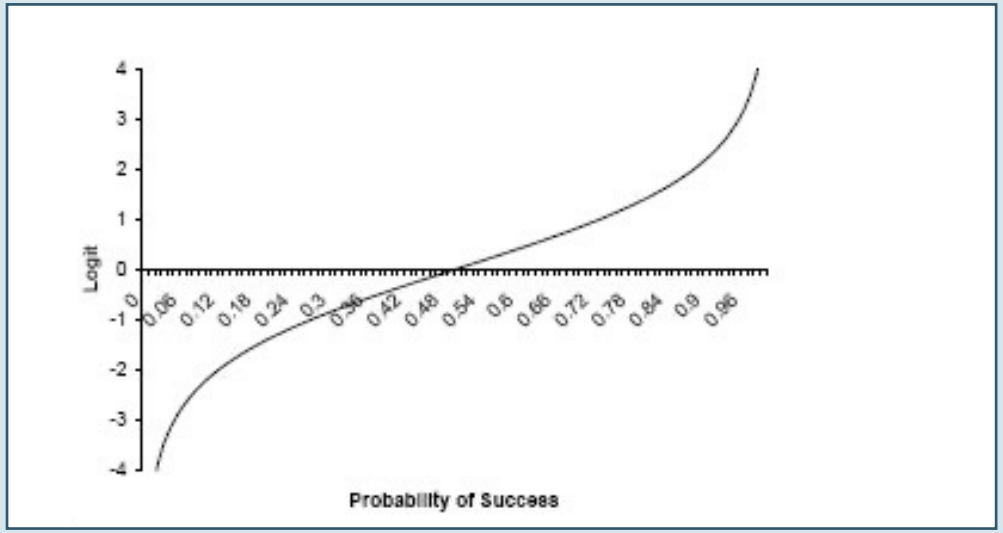
Assumptions

- The only “real” limitation on logistic regression is that the outcome must be discrete.
- Ratio of cases to variables – using discrete variables requires that there are enough responses in every given category
  - If there are too many cells with no responses parameter estimates and standard errors will likely blow up
  - Also can make groups perfectly separable (e.g. multicollinear) which will make maximum likelihood estimation impossible.
- Linearity in the logit – the regression equation should have a linear relationship with the logit form of the discrete variable. There is no assumption about the predictors being linearly related to each other.

The Logit

- Goal: Find a function of the predictor variables that relates them to a 0/1 outcome
- Instead of Y as outcome variable (like in linear regression), we use a function of Y called the **logit**
- Logit can be modeled as a linear function of the predictors
- The logit can be mapped back to a probability, which, in turn, can be mapped to a class

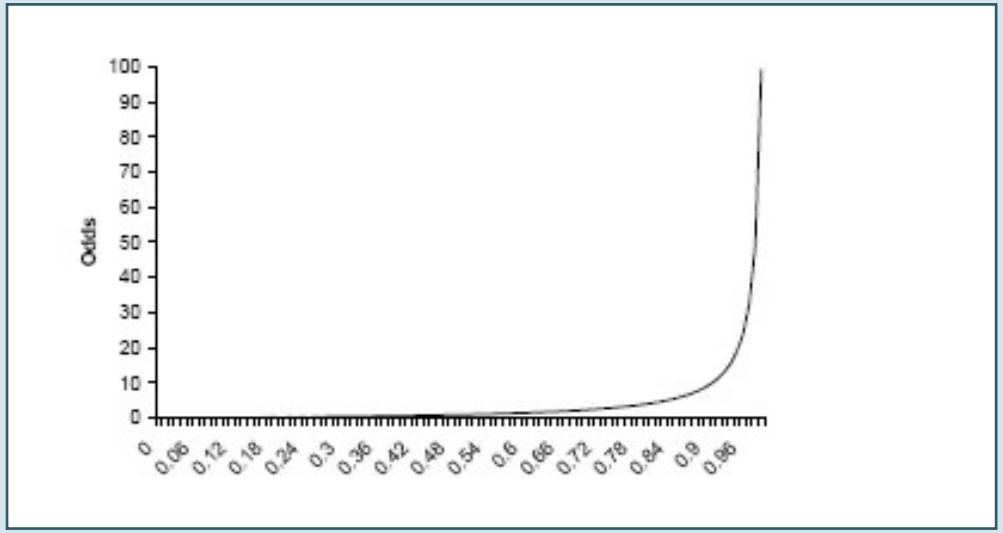
Example: Logit as a function of p



We now define the odds and how they can be computed

- The odds of an event are defined as:
  - $Odds = \frac{p}{1-p}$ ← p = probability of event
- Or, given the odds of an event, the probability of the event can be computed by:
  - $p = \frac{Odds}{1+Odds}$
- We can also relate the Odds to the predictors:
  - $Odds = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+ \ldots \beta_q x_q}$

Example: Odds as a function of p



First we define the Logistic Loss Function

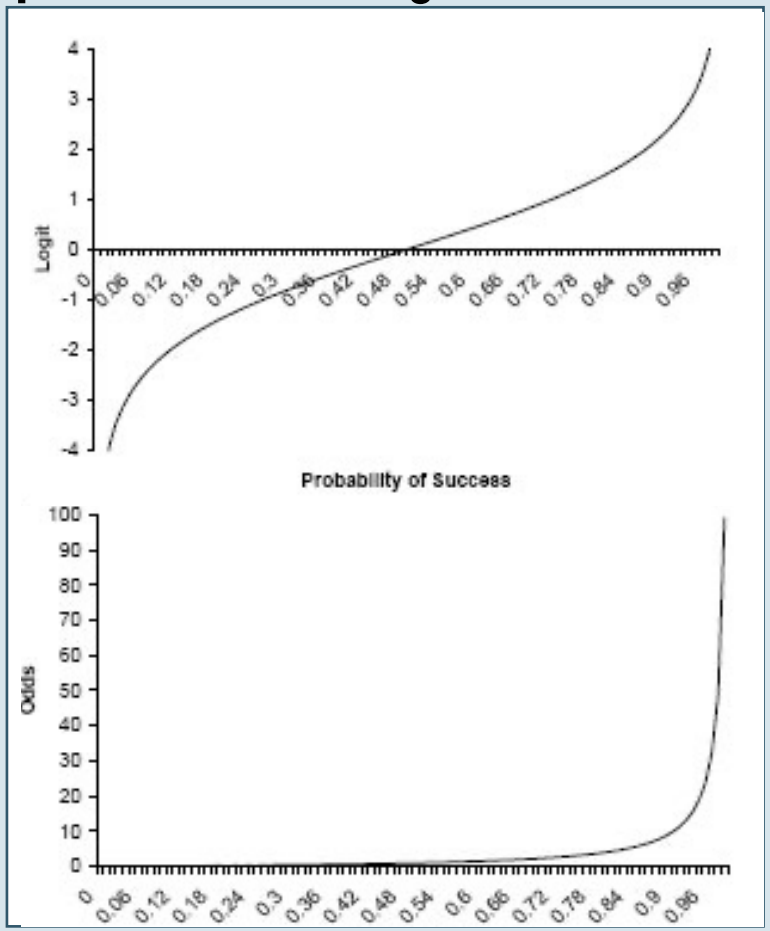
- p = probability of belonging to class 1
- Need to relate p to predictors with a function that guarantees $0 \leq p \leq 1$
- Standard linear function (as shown below) does not:
  - $p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \beta_q x_q$

- The Fix: use **logistic loss function**
  - $p = \frac{Odds}{1+Odds}$
  - $p = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+ \ldots \beta_q x_q)}}$

Take log on both sides

- This gives us the logit:
  - $\log(Odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_q x_q$
  - $\log(Odds) = logit$
- So, the logit is a linear function of predictors $x_1, x_2, \ldots$
  - Takes values from $-\infty$ to $+\infty$

Example: Odds and logit as a function of p



Recovering Probabilities

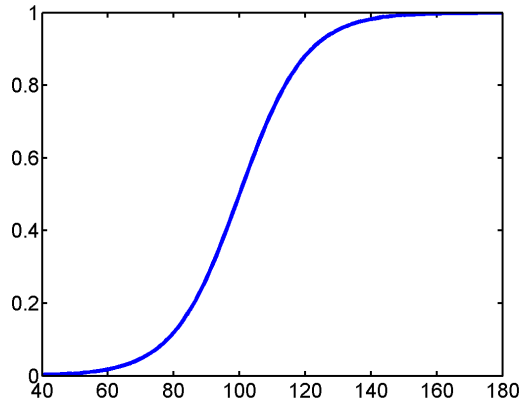
- $ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$

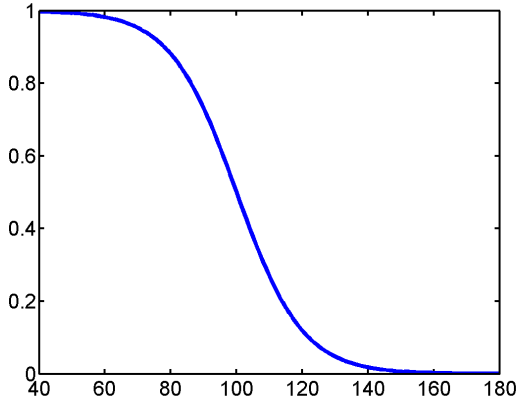
$$\Leftrightarrow \frac{p}{1-p} = e^{\beta_0+\beta_1 X}$$

$$\Leftrightarrow p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

- which gives p as a sigmoid function!

- **Logistic Response Function**
  - When the response variable is binary, the shape of the response function is often sigmoidal:





Interpretation of β_1

- Let:
  - odds1 = odds for value X ($\frac{p}{1-p}$)
  - odds2 = odds for value $X + 1$ unit
- Then:
  - $\frac{odds2}{odds1} = \frac{e^{\beta_0+\beta_1(X+1)}}{e^{\beta_0+\beta_1 X}} = \frac{e^{(\beta_0+\beta_1 X)+\beta_1}}{e^{\beta_0+\beta_1 X}} = \frac{e^{(\beta_0+\beta_1 X)}e^{\beta_1}}{e^{\beta_0+\beta_1 X}} = e^{\beta_1}$
- Hence, the exponent of the slope describes the proportionate rate at which the predicted odds ratio changes with each successive unit of X

Sample Calculations

- Suppose a cancer study yields: log odds = –2.6837 + 0.0812 SurvRate

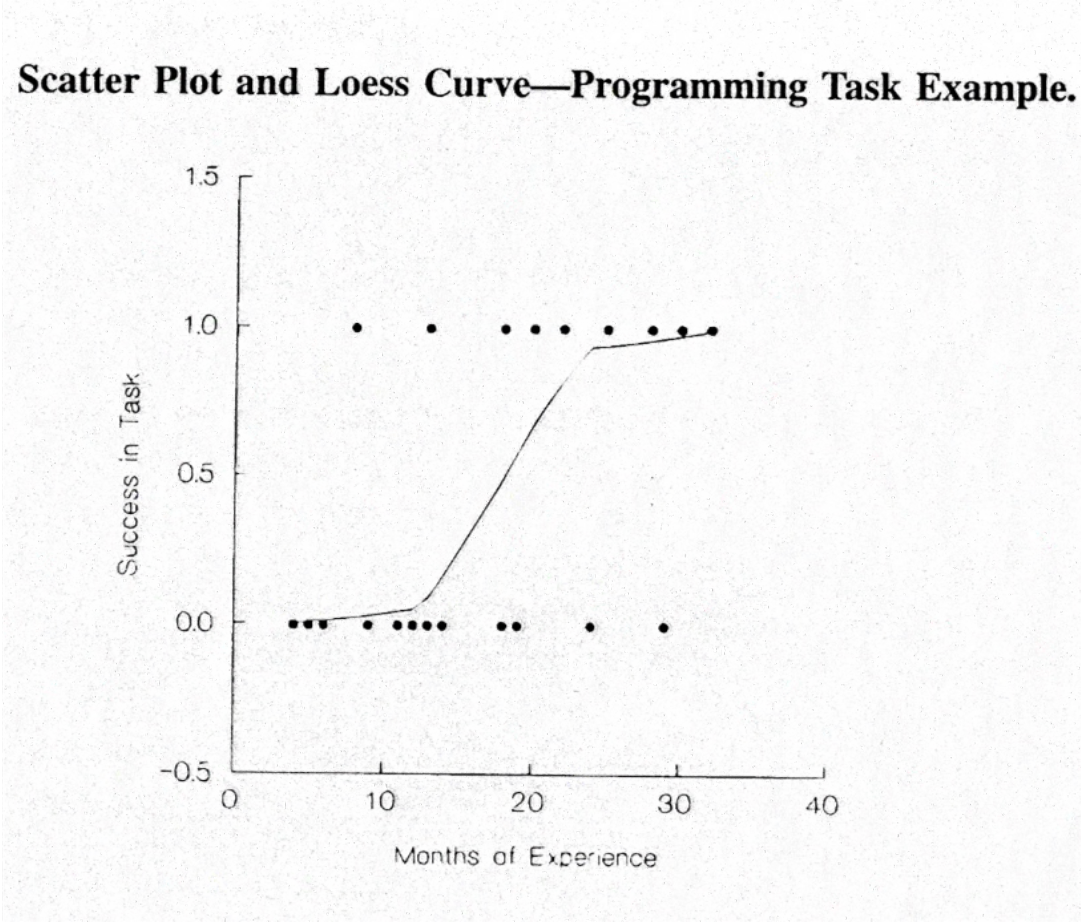
- Consider a patient with SurvRate = 40
  - log odds = –2.6837 + 0.0812(40) = 0.5643
  - odds = e0.5643 = 1.758
  - patient is 1.758 times more likely to be improved than not

- Consider another patient with SurvRate = 41
  - log odds = –2.6837 + 0.0812(41) = 0.6455
  - odds = e0.6455 = 1.907
  - patient's odds are 1.907/1.758 = 1.0846 times (or 8.5%) better than those of the previous patient

- Using probabilities
  - P(40) = 0.6374 and P(41) = 0.6560
  - Improvements appear different with odds and with p

Example 1 (I)

- A systems analyst studied the effect of computer programming experience on ability to complete a task within a specified time
- Twenty-five persons selected for the study, with varying amounts of computer experience (in months)
- Results are coded in binary fashion: Y = 1 if task completed successfully; Y = 0, otherwise



- Loess: form of local regression

Example 1 (II)

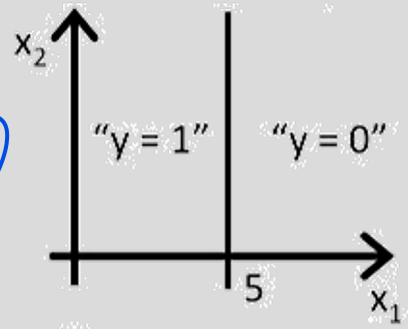
- Results from a standard package give:
  - $\beta_0 = -3.0597 \; and \; \beta_1 = 0.1615$
- Estimated logistic regression function:
  - $p = \frac{1}{1+e^{3.0597-0.1615X}}$
- For example, the fitted value for $X = 14$ is:
  - $p = \frac{1}{1+e^{3.0597-0.1615(14)}} = 0.31$
  - (Estimated probability that a person with 14 months experience will successfully complete the task)

Example 1 (III+IV)

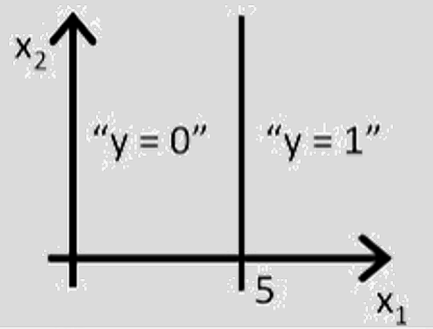
- We know that the probability of success increases sharply with experience
  - Odds ratio: $exp(\beta_1) = e^{0.1615} = 1.175$
  - Odds increase by 17.5% with each additional month of experience
- A unit increase of one month is quite small, and we might want to know the change in odds for a longer difference in time
  - For c units of X : $exp(c\beta_1)$
- Suppose we want to compare individuals with relatively little experience to those with extensive experience, say 10 months versus 25 months $(c = 15)$
  - Odds ratio: $e^{15 \times 0.1615} = 11.3$
  - Odds of completing the task increase 11-fold!

Consider logistic regression with two features x_1 and x_2. Suppose $\theta_0 = 5, \theta_1 = -1, \theta_2 = 0$, so that $h(x)_\theta = g(5 - x_1)$. Which of these shows the decision boundary of $h(x)_\theta$?

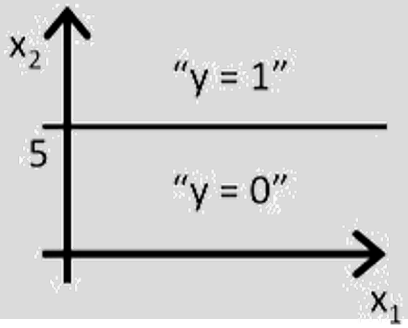
a)



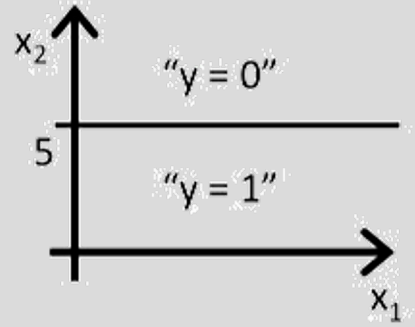
b)



c)

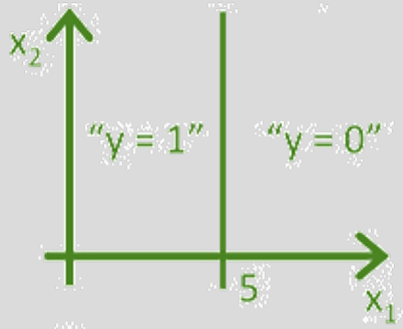


d)

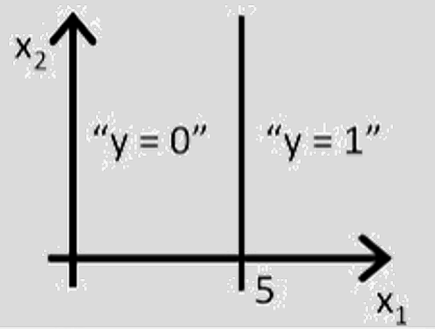


Consider logistic regression with two features x_1 and x_2. Suppose $\theta_0 = 5, \theta_1 = -1, \theta_2 = 0$, so that $h(x)_\theta = g(5 - x_1)$. Which of these shows the decision boundary of $h(x)_\theta$?

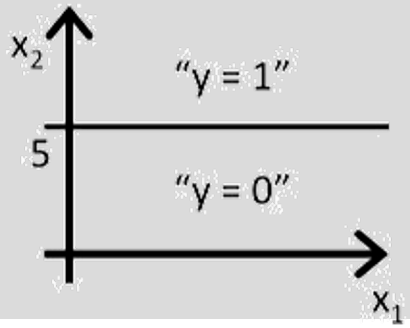
a)



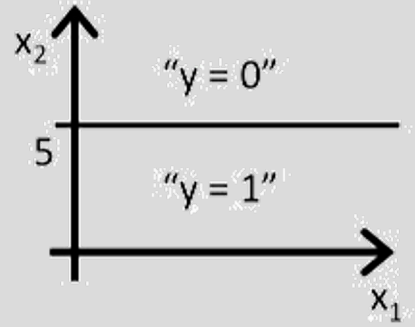
b)



c)



d)



Predict Y = 0 if x_1 is greater than 5

Summary

- Logistic regression is similar to linear regression, except that it is used with a categorical response
- It can be used for explanatory tasks (=profiling) or predictive tasks (=classification)
- The predictors are related to the response Y via a nonlinear function called the **logit**
- As in linear regression, reducing predictors can be done via variable selection
- Logistic regression can be generalized to more than two classes

Agenda

- Linear Classifiers (continue)
- Classification via Logistic Regression
- **Review probability**
- Naïve Bayes Classifier

Need for Probabilistic Reasoning

e.g. election forecasts

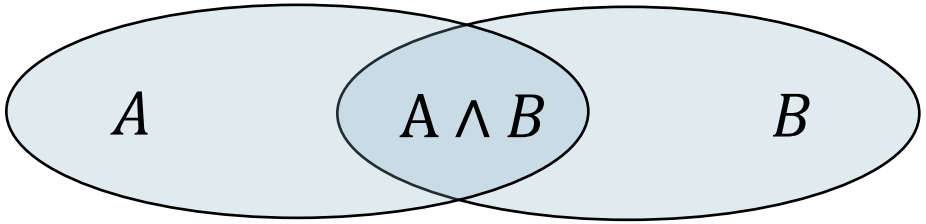
1. Most everyday reasoning is based on **uncertain evidence and inferences**
2. **Classical logic**, which only allows conclusions to be strictly true or strictly false, **does not account** for this **uncertainty** or the need to weigh and combine conflicting evidence
3. **Straightforward application of probability theory** is **impractical** since the large number of probability parameters required are rarely, if ever, available
4. Therefore, early expert systems employed fairly ad hoc methods for reasoning under uncertainty and for combining evidence
5. **Recently**, **methods more rigorously founded in probability theory** that attempt to decrease the amount of conditional probabilities required have flourished (e.g. Naïve Bayes Classifiers)

Notation for random variables

- In this lecture, we use upper case letters, X to denote random variables
- For a random variable X taking values {1, 2, 3}:
  - $p(X) = \begin{cases} 1 : 0.1 \\ 2 : 0.5 \\ 3 : 0.4 \end{cases}$ represents a mapping from values to probabilities numbers that sum to one (odd notation, would be better to use p_X, but this is not common)
- Conversely, we will use lower case x to denote a specific value of X (i.e., for above example $x \in \{1,2,3\}$, and $p(X = x)$ or just $p(x)$ refers to a number (the corresponding entry of $p(X)$)

Axioms of Probability Theory

- All probabilities between 0 and 1
  - $0 \leq P(A) \leq 1$
- True proposition has probability 1, false has probability 0
  - P(true) = 1
  - P(false) = 0.
- The joint probability is: $A \wedge B$
- The probability of disjunction is: $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

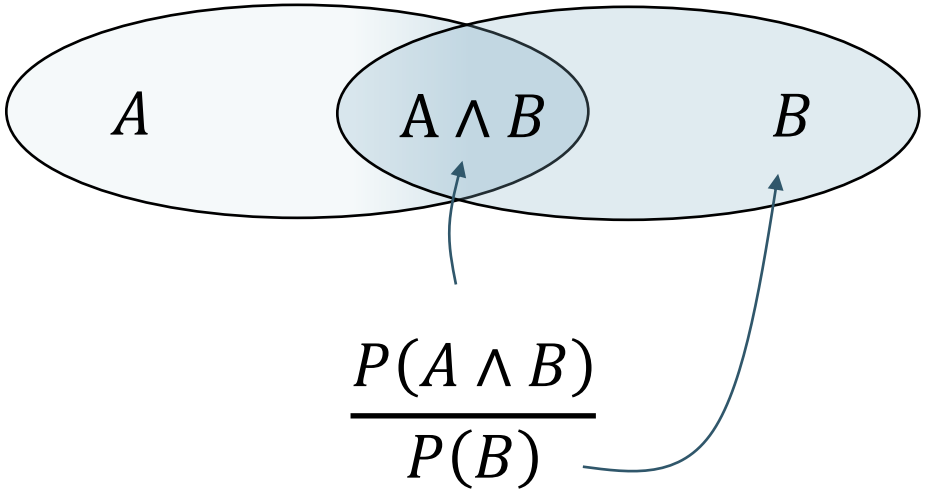


Conditional Probability – Probability of seeing A given that we have observed B

- $P(A|B)$ is the probability of A given B
- Assumes that B is all and only information known/observed
- Defined by: $P(A|B) = \frac{P(A \wedge B)}{P(B)}$

joint probability

e.g. prob. of buying item if price = 60 €



Independence – Observing A does not tell you anything about B

- A and B are independent if:
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$
  - These two constraints are logically equivalent

- Therefore, if A and B are independent:
  - $P(A|B) = \frac{P(A \wedge B)}{P(B)} = P(A)$
  - $P(A \wedge B) = P(A)P(B)$

Joint probability – What is the probability of A and B occurring together?

- The joint probability distribution for a set of random variables, $X_1,\ldots,X_n$ gives the probability of every combination of values (an *n*-dimensional array with v^n values if all variables are discrete with *v* values, all v^n values must sum to 1): $P(X_1,\ldots,X_n)$

positive

	circle	square
red	0.20	0.02
blue	0.02	0.01

negative

	circle	square
red	0.05	0.30
blue	0.20	0.20

- The probability of all possible conjunctions (assignments of values to some subset of variables) can be calculated by summing the appropriate subset of values from the joint distribution

Joint probability – Example calculations

- Task: Calculate $\mathrm{P}(red \wedge circle)$ and $P(red)$:
  - $P(red \wedge circle) = 0.20 + 0.05 = 0.25$
  - $P(red) = 0.20 + 0.02 + 0.05 + 0.3 = 0.57$
- Therefore, all conditional probabilities can also be calculated
- Task: Calculate $P(positive \mid red \wedge circle)$:

$$= \frac{P(positive \wedge red \wedge circle)}{P(red \wedge circle)}$$

$$= \frac{0.20}{0.20 + 0.05} = 0.80$$

positive

	circle	square
red	0.20	0.02
blue	0.02	0.01

negative

	circle	square
red	0.05	0.30
blue	0.20	0.20

Joint probability – Deep-dive on notation

- Given two random variables: X with values in {1,2,3} and Y with values in {1,2} :
  - $\mathrm{p}(\mathrm{X}, Y)$ refers to the **joint probability distribution**, i.e., a set of 6 possible values for each setting of variables, i.e. a dictionary mapping $(1,1)$, $(1,2)$, $(2,1)$, ... to corresponding probabilities)
  - $\mathrm{p}(\mathrm{x}_1, \mathrm{y}_2)$ is a number: probability that $X = \mathrm{x}_1$ and $Y = \mathrm{y}_2$
  - $\mathrm{p}(\mathrm{X}, \mathrm{y}_2)$ is a set of 3 values, the probabilities for all values of X for the given value $Y = \mathrm{y}_2$, i.e., it is a **dictionary mapping** 1,2,3 to 2 (note: not a probability distribution, it will not sum to one)
- We generally call all of these terms factors (dictionaries mapping values to numbers, even if they do not sum to one)

Joint probability – Example of weather and cavity

$$P(Weather,\ Cavity) = \begin{cases} sunny, yes\ 0.07 \\ sunny, no\ 0.63 \\ rainy, yes\ 0.02 \\ rainy, no\ 0.18 \\ cloudy, yes\ 0.01 \\ cloudy, no\ 0.09 \end{cases}$$

- Let Weather denote a random variable taking on values in **Weather = {sunny, rainy, cloudy}**
- Let Cavity denote a random variables taking on values in **Cavity = {yes, no}**

Joint probability – p(sunny, yes)?

$$P(Weather,\ Cavity) = \begin{cases} sunny, yes\ \ 0.07 \\ sunny, no\ \ 0.63 \\ rainy, yes\ \ 0.02 \\ rainy, no\ \ 0.18 \\ cloudy, yes\ \ 0.01 \\ cloudy, no\ \ 0.09 \end{cases}$$

$$\Rightarrow\ p(sunny, yes) = 0.07$$

- Let Weather denote a random variable taking on values in **Weather = {sunny, rainy, cloudy}**
- Let Cavity denote a random variables taking on values in **Cavity = {yes, no}**

Joint probability – p(Weather, yes)?

$$P(Weather,\ Cavity) = \begin{cases} sunny, yes\ 0.07 \\ sunny, no\ 0.63 \\ rainy, yes\ 0.02 \\ rainy, no\ 0.18 \\ cloudy, yes\ 0.01 \\ cloudy, no\ 0.09 \end{cases}$$

$$\Rightarrow\ p(Weather, yes) = \begin{cases} sunny\ 0.07 \\ rainy\ 0.02 \\ cloudy\ 0.01 \end{cases}$$

- Let Weather denote a random variable taking on values in **Weather = {sunny, rainy, cloudy}**
- Let Cavity denote a random variables taking on values in **Cavity = {yes, no}**

Joint probability – Another notation of joint probability

- The **conditional probability** $p(X_1|X_2)$ (the conditional probability of X_1 given X_2) is defined as
  - $p(X_1|X_2) = \frac{p(X_1,X_2)}{p(X_2)}$
- Therefore, the **joint probability** can **also be written** $\mathbf{p(X_1, X_2) = p(X_1|X_2)p(X_2)}$

Marginalization

- For random variables $\mathrm{X}_1, \mathrm{X}_2$ with joint distribution $\mathrm{p}(\mathrm{X}_1, \mathrm{X}_2)$
  - $p(X_1) = \sum_{x_2} p(X_1, x_2) = \sum_{x_2} p(X_1|x_2)p(x_2)$
- Generalizes to joint distributions over multiple random variables
  - $p(X_1, \dots, X_i) = \sum_{x_{i+1},\dots,x_n} p(X_1, \dots, X_i, x_{i+1}, \dots, x_n)$
- For p to be a probability distribution, the marginalization over all variables must be one
  - $\sum_{x_1,\dots,x_n} p(x_1, \dots, x_n) = 1$

Independence revisited

- We say that random variables X_1 and X_2 are (marginally) independent if their joint probability distribution is the product of their marginal distributions
  - $p(X_1, X_2) = p(X_1)p(X_2)$

- Equivalently, can also be stated as the condition that
  - $p(X_1|X_2)\left(= \frac{p(X_1,X_2)}{p(X_2)} = \frac{p(X_1)p(X_2)}{p(X_2)}\right) = p(X_1)$
  - **(and similarly)** $\boldsymbol{p(X_2|X_1) = p(X_2)}$

Are the Weather and Cavity random variables independent? In other words, does the following hold true p$(Weather|Cavity) = p(Weather)$?

a) Yes
b) No

$$P(Weather,\ Cavity) = \begin{cases} sunny, yes\ \ 0.07 \\ sunny, no\ \ 0.63 \\ rainy, yes\ \ 0.02 \\ rainy, no\ \ 0.18 \\ cloudy, yes\ \ 0.01 \\ cloudy, no\ \ 0.09 \end{cases}$$

Are the Weather and Cavity random variables independent? In other words does the following hold true p$(Weather|Cavity) = p(Weather)$?

a) Yes
b) No

$$P(Weather,\ Cavity) = \begin{cases} sunny, yes\ \ 0.07 \\ sunny, no\ \ 0.63 \\ rainy, yes\ \ 0.02 \\ rainy, no\ \ 0.18 \\ cloudy, yes\ \ 0.01 \\ cloudy, no\ \ 0.09 \end{cases}$$

$$p(Cavity = Yes) = 0.07 + 0.02 + 0.01 = 0.1$$

$$p(Weather) = \begin{cases} p(Weather = Sunny) = 0.07 + 0.63 = 0.7 \\ p(Weather = Rainy) = 0.02 + 0.18 = 0.2 \\ p(Weather = Cloudy) = 0.01 + 0.09 = 0.1 \end{cases}$$

$$p(Weather \,|Cavity = Yes)$$

$$= \begin{cases} p(Weather = Sunny \,|Cavity = Yes) = \dfrac{0.07}{0.1} = 0.7 \\ p(Weather = Rainy \,|Cavity = Yes) = \dfrac{0.02}{0.1} = 0.2 \\ p(Weather = Cloudy \,|Cavity = Yes) = \dfrac{0.01}{0.1} = 0.1 \end{cases}$$

Bayes's rule – A straightforward manipulation of probabilities using the rules derived before

Bayes's rule

Assuming X_1 and X_2 are independent

- $p(X_1|X_2) = \frac{p(X_1, X_2)}{\mathrm{p}(X_2)}$

$= \frac{p(X_2|X_1)p(X_1)}{\mathrm{p}(X_2)}$

As per the definition of joint probability:

$P(X_1|X_2) = P(X_1,X_2) / P(X_2)$

$= \frac{p(X_2|X_1)p(X_1)}{\sum_{x_1} p(X_2|x_1)p(x_1)}$

Marginalize out x_1

Intuition

- Relate the actual probability to the measured test probability
- The quantities $p(X_2|X_1)$, $p(X_1)$, and $\mathrm{p}(X_2)$ may be easier to determine than the quantity of ultimate interest: $p(X_1|X_2)$
- Combined with the assumption of independence, the Bayes' rule of conditional probability serves as basis for the Naïve Bayes Classifier that we will discuss later on

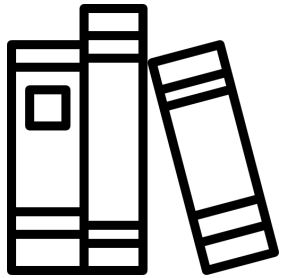
Bayes's rule – An intuitive example based on Kahnemann and Tversky`s "Librarian or Farmer" question

Description of Steve (a random person)

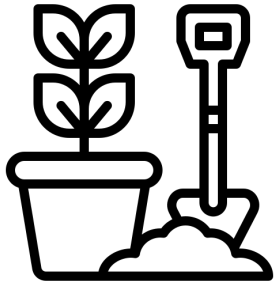
*"Steve is **very shy and withdrawn**, invariably helpful but with little interest in people or in the world of reality. A **meek and tidy soul**, he has a need for order and structure, and a passion for detail."*

What is Steve's job?

Steve is a **librarian**



Steve is a **farmer**

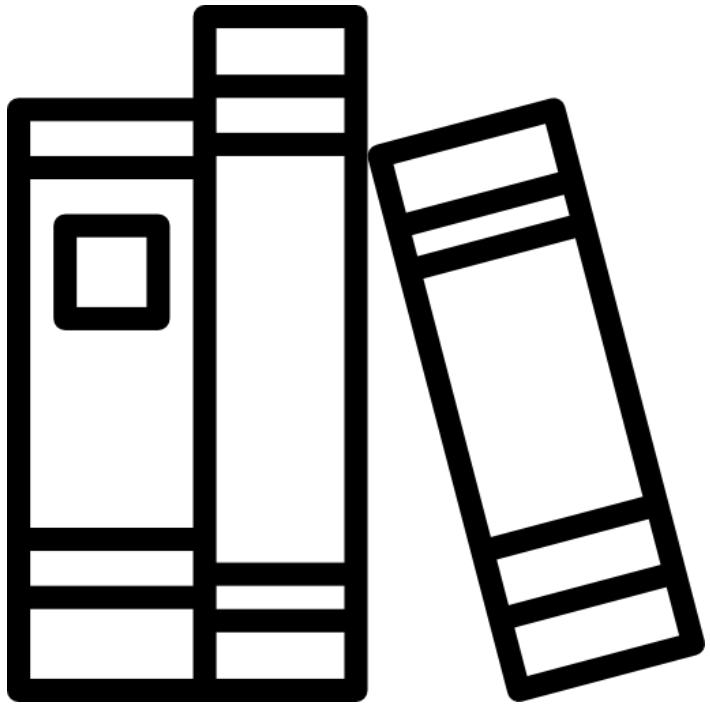


What do you think? Given the description of traits, is Steve more likely to be

a) a librarian,
or
b) a farmer?

Bayes's rule – Most people would say that Steve is more likely to be a librarian

Steve is a **librarian!**



Circa. 90%

Steve is a **farmer!**

Circa. 10%

Bayes's rule – But Bayes's rule tells us differently. Let's plug in the numbers!

Goal and Data

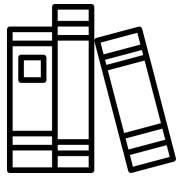
- Our goal is to **estimate**
  - P(Job = Librarian | Traits = Description)
- To do so we need to **collect some more information (also known as priors)** – Let's assume the following (**our prior beliefs**):
  - A ratio of librarians to farmers of 1:20
  - 40% of Librarians fit the description of "meek and tidy"
  - 10% of farmers fit the description of "meek and tidy"

Calculations

- According to Bayes's rule, we have the following relationship:
  - P(Librarian | Description) = (P(Description | Librarian)* P(Librarian)) / P(Description)
- We can **compute** all **required prior beliefs**
  - P(Job=Librarian) = 1/21 = 0.0476
  - P (Traits=Description | Job=Librarian) = 0.4
  - P(Traits = Description) = (0.4*1 + 0.1*20)/21 = 0.1142
- Subbing this into the above equations yields:
  - (0.4*0.0476)/0.12 = 0.1667, i.e., approx. 17%

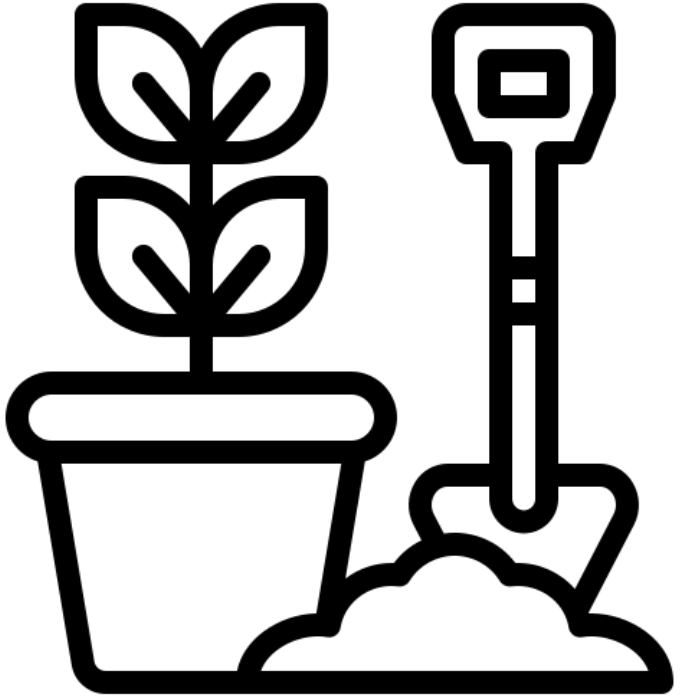
Bayes's rule – So using the laws of probability and some well-grounded assumptions (prior beliefs), people should have said the following

Steve is a **librarian!**



Circa. 17%

Steve is a **farmer!**



Circa. 83%

Bayes's rule – Still a bit confused? Watch this excellent video for a graphical intuition using the same example and numbers!

- Bayes theorem by 3Bule1Brown, accessible here

I want to know if I have come down with a rare strain of flu (occurring in only 1/ 10,000 people). There is an “accurate” test for the flu (if I have the flu, it will tell me I have it 99% of the time, and if I do not have it, it will tell me I do not have it 99% of the time). I go to the doctor and test positive.

What is the probability I have the flu?

Answer: Bayes' rule

X_1: Have flu ($X_1 = \{0,1\}$); X_2 : Test positive ($X_2 = \{0,1\}$)

Prior beliefs:

$$p(X_2 = 1|X_1 = 1) = 0.99\ ; p(X_1 = 1) = \frac{1}{10000} = 0.0001$$

$$p(X_1 = 1|X_2 = 1)$$

$$= \frac{p(X_2 = 1|X_1 = 1)p(X_1 = 1)}{p(X_2 = 1|X_1 = 0)p(X_1 = 0) + p(X_2 = 1|X_1 = 1)p(X_1 = 1)}$$

$$= \frac{0.99 \times 0.0001}{0.01 \times 0.9999 + 0.99 \times 0.0001} \approx \mathbf{0.01}$$

Contact

For general questions and enquiries on **research**, **teaching**, **job openings** and new **projects** refer to our website at www.is3.uni-koeln.de

For specific enquiries regarding this course contact us by sending an email to the **IS3 teaching** address at is3-teaching@wiso.uni-koeln.de