Logistic Regression

COS 424/524, SML 302: Fundamentals of Machine Learning

Professor Engelhardt

COS424/524, SML 302

Lecture 9

Classification using regression

In previous lectures, we learned how to perform:

- classification with generative models
- classification with discriminative models
- prediction with linear regression

In this lecture, we develop linear models that act as discriminative classifiers.

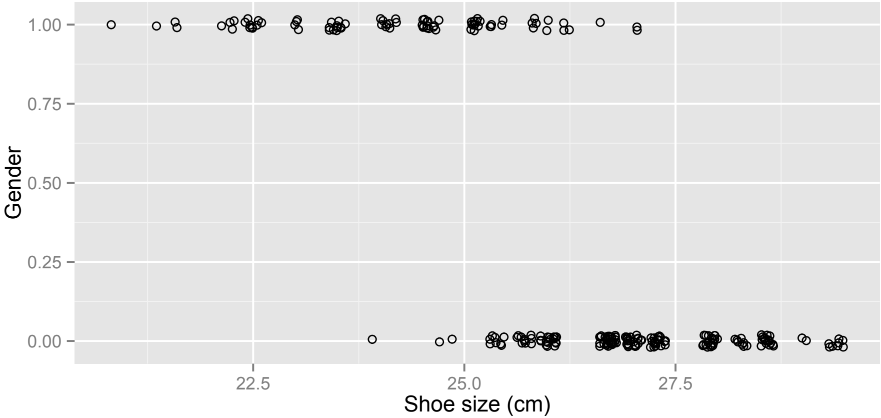
Recall: the problem of classification

Classification examples

- Classify an email as “spam” or “not spam” from text
- Classify images into categories: “cat” or “beach” or “typewriter”
- Classify news articles into newspaper sections: “politics” or “sports”
- Classify genetic code as “exon” or “intron”
- Classify radar blips as “friendly” or “unfriendly”
- Classify credit cards as “stolen” or “not stolen” from activity
- Classify patient as “has disease” or "healthy” from medical record
- Many others...

Classification: example

Let's plot one example of this problem (with jitter): Can we classify a person by gender by looking at shoe size?



How can we use linear regression to predict values near zero and one?

Regression for classification

Linear regression corresponds to a discriminative graphical model:
$p(x_i, y_i) \propto p(y_i \mid x_i)$

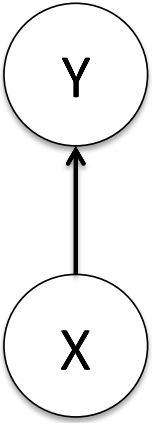
We can use regression to perform classification.

Let's consider binary classification, where each data point is in one of two classes $y_i \in \{0, 1\}$.

If we used linear regression to model these data, then

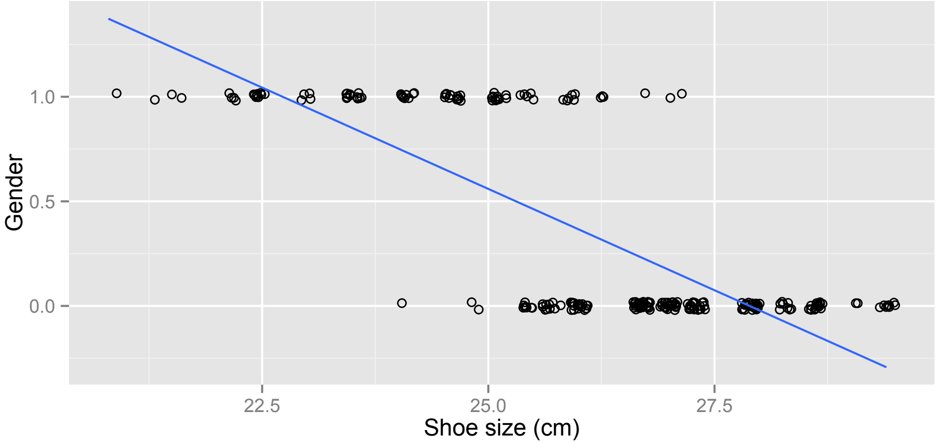
$$y_i \mid x_i \sim \mathcal{N}(\beta^\top x_i, \sigma^2).$$

Is linear regression appropriate for binary classification?



Classification: example

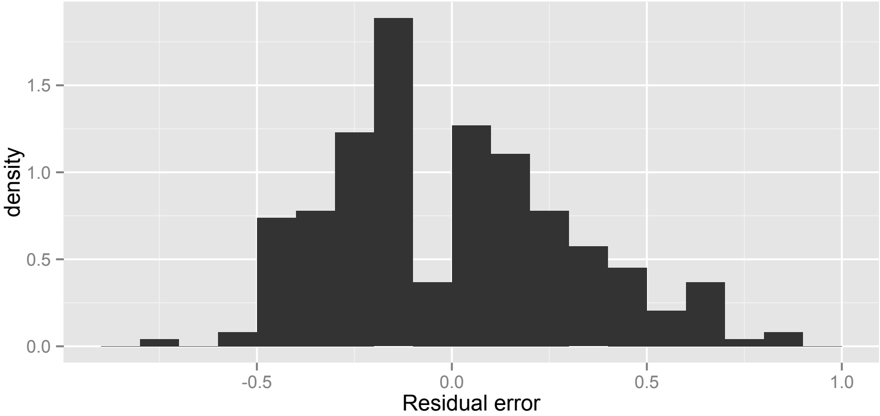
Can we classify a person by gender using shoe size with linear regression?



Is this a good classifier?

Classification: example

If we classify a person by gender by looking at their shoes using linear regression, what do the residuals look like?



While linear regression is reasonable, residual $y - \hat{y}$ non-Gaussian, incorrect

Bernoulli response model

Try direct approach: model conditional distribution of y, $p(y = 1 \mid x)$, explicitly as a Bernoulli whose bias parameter is a function of x:

$$
\begin{aligned}
p(y \mid x) &= \mu(x)^y (1 - \mu(x))^{1-y} \\
p(y = 1 \mid x) &= \mu(x).
\end{aligned}
$$

What form should $\mu(x)$ take?

Bernoulli response model

We model conditional distribution of y, $p(y = 1 \mid x)$, as a Bernoulli whose bias parameter is a function of x:

$$p(y = 1 \mid x) \quad = \quad \mathrm{E}[y \mid x] = \mu(x).$$

Let's go back to our line of thinking around linear regression.

Linear regression is Gaussian with mean a function of x, specifically

$$\begin{aligned} y &\sim \mathcal{N}(\mu(x), \sigma^2) \\ \mu(x) &= \beta^\top x \end{aligned}$$

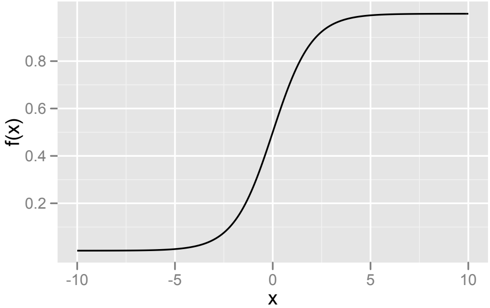
Is this appropriate for the Bernoulli?

Logistic function

To model the bias parameter, we use the *logistic function*,

$$\mu(x) = \frac{1}{1 + e^{-\eta(x)}}.$$

This function maps $x \in \Re$ to a value in $(0, 1)$.



What happens when $\eta(x) = -\infty, +\infty, 0$?

In logistic regression, as in linear regression, we set

$$\eta(x) = \beta^\top x.$$

and choose $\mu(\cdot)$ to be the logistic function. This specifies the model,

$$\begin{aligned} y_i \mid x_i, \beta &\sim \text{Bernoulli}(\mu(\beta^\top x_i)) \\ &= \text{Bernoulli}\left(\frac{1}{1 + e^{-\beta^\top x_i}}\right) \\ p(y_i = 1 \mid x_i, \beta) &= \frac{1}{1 + e^{-\beta^\top x_i}}. \end{aligned}$$

What is the role of the intercept term here?

Logistic regression

Model for logistic regression:

$$y_i | x_i, \beta \quad \sim \quad \text{Bernoulli}(\mu(\beta^\top x_i)).$$

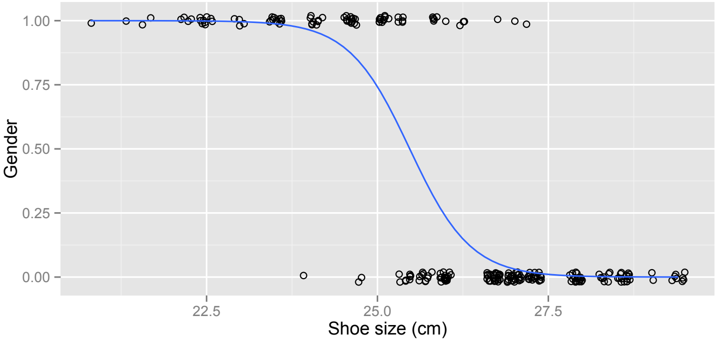
Key point: *The covariates enter the probability of the response through a linear combination with the coefficients.*

That linear combination is then passed through function μ to be appropriate as a parameter for the distribution of the response.

Can this be generalized to other response conditional distributions?

Classification: example

Can we classify a person by gender using shoe size with logistic regression?

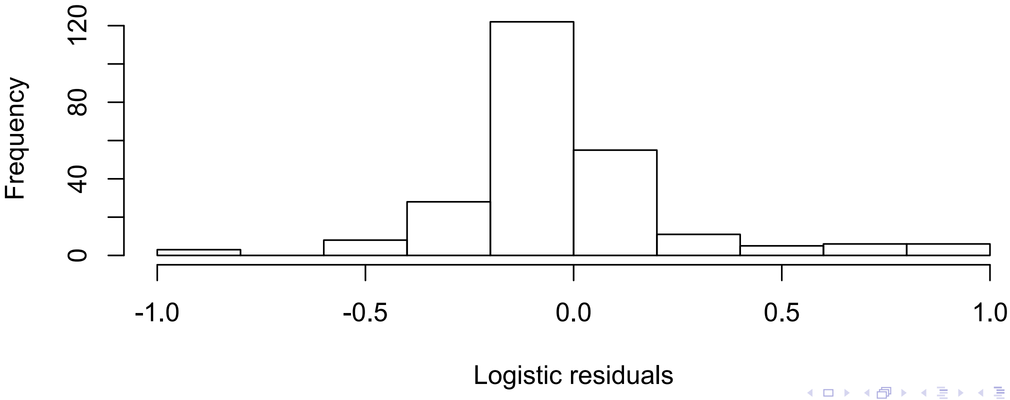


Is this a good classifier?

Residual sum of squares (logistic model): 17.4 Residual sum of squares (linear model): 23.3

Classification: example

If we classify a person by gender by looking at their shoes using logistic regression, what do the residuals look like?



Logistic regression: intercept

As with linear regression, we generally include an intercept term β_0:

$$\mathrm{E}[y_i | x_i] \quad = \quad \frac{1}{1 + e^{-(\beta_0 + \beta x_i)}}$$

The intercept term is an offset on the x axis for the logistic function: when $\beta_0 = -\beta x_i$, this is where

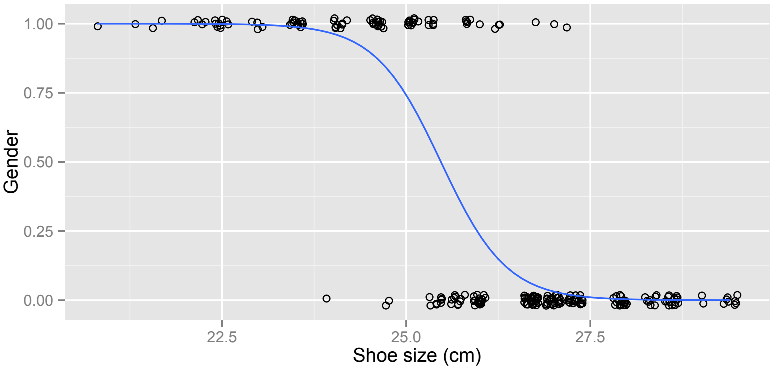
$$p(y_i = 1 \mid x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta x_i)}} = \frac{1}{2},$$

or

$$(\beta_0 + \beta x_i) = 0.$$

Classification: example

Can we classify a person by gender using shoe size with logistic regression?



When $(\beta_0 + \beta x_i) = 0$, and $p(y_i = 1 \mid x_i) = \frac{1}{2}$, intercept $-\beta_0 = \beta x_i$. Estimated intercept, slope (logistic model): $\beta_0 = 57.1$, $\beta = -2.24$

Logistic regression: coefficients

As with linear regression, interpret estimates of coefficients β:

$$p(y_i = 1|x_i) = \frac{1}{1 + e^{-\beta^\top x_i}}$$

Let us rewrite this equation in terms of β:

$$\begin{aligned}
\frac{1}{p(y_i = 1|x_i)} &= 1 + e^{-\beta^\top x_i} \\
\frac{1}{p(y_i = 1|x_i)} - 1 &= e^{-\beta^\top x_i} \\
\frac{1 - p(y_i = 1|x_i)}{p(y_i = 1|x_i)} &= e^{-\beta^\top x_i} \\
\log \frac{1 - p(y_i = 1|x_i)}{p(y_i = 1|x_i)} &= -\beta^\top x_i \\
\log \frac{p(y_i = 1|x_i)}{1 - p(y_i = 1|x_i)} &= \beta^\top x_i
\end{aligned}$$

Logistic regression: coefficients

Let's look at this equation:

$$\log \frac{p(y_i = 1|x_i)}{1 - p(y_i = 1|x_i)} \quad = \quad \beta^\top x_i$$

First, what is the term: $\frac{p(y_i=1|x_i)}{1-p(y_i=1|x_i)} = \frac{p(y_i=1|x_i)}{p(y_i=0|x_i)}$? It is an *odds ratio*: the ratio of the probability of success over the probability of failure.

Odds ratio example

If the probability of a coin coming up heads is 0.7, then the odds of a head for one coin flip is:

$$\frac{0.7}{1-0.7} = \frac{0.7}{0.3} = 2.333$$

Detour: odds and the odds ratio

How do *odds* relate to the *odds ratio*?

In gambling, odds are most often described as *odds against* a success.

Odds X to Y tells us that, out of $X + Y$ total events, in expectation Y will be successful, X will be failure.

Odds examples

- The odds against a fair die coming up a six is 5 to 1.
- The odds in favor of the Kansas City Chiefs beating the Tampa Bay Buccaneers is 6 to 1.
- The odds against *The Shape of Water* winning best picture at the 2019 Oscars was 13 to 8.
- The odds against *Lady Bird* winning best picture at the 2019 Oscars was 12 to 1.

What sum, across Oscar movies, must be one for odds to have a probabilistic interpretation?

Logistic regression: coefficients

Returning to this equation:

$$\log \frac{p(y_i = 1|x_i)}{1 - p(y_i = 1|x_i)} \quad = \quad \beta^\top x_i$$

The product of β and x_i is the log odds ratio: $\log \frac{p(y_i=1|x_i)}{1-p(y_i=1|x_i)}$.

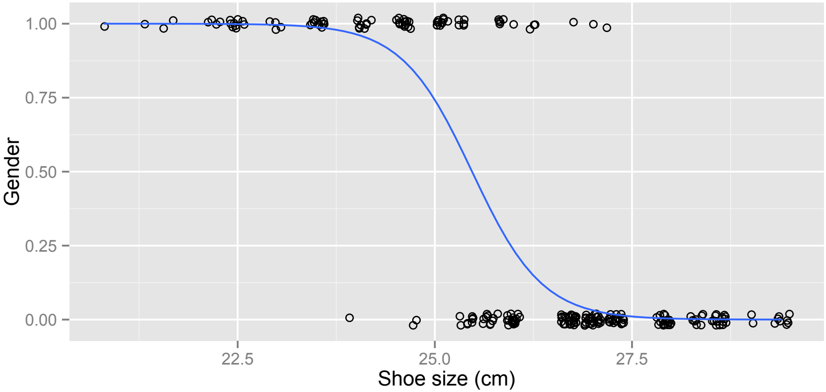
This means that:

- when β is positive, relationship of x and y will be proportional.
- when β is negative, relationship of x and y will be inversely proportional.
- when β is zero, as in linear regression, x is not predictive of y.

Can we use regularization for logistic regression?

Classification: example

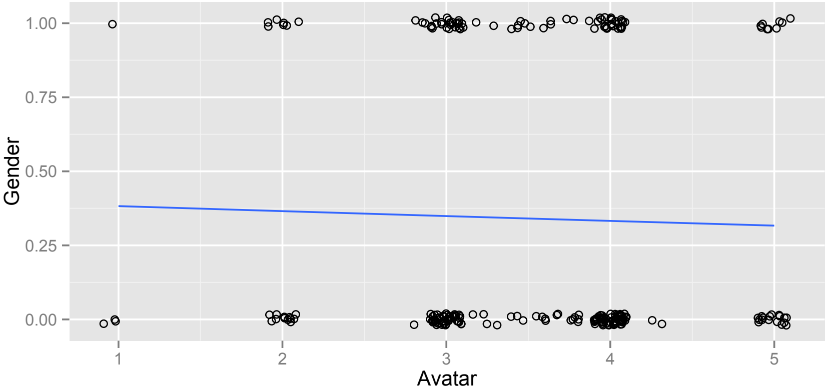
Can we classify a person by gender using shoe size with logistic regression?



Residual sum of squares (logistic model): 17.4
Residual sum of squares (linear model): 23.3

Classification: example

Can we classify self-reported gender using *Avatar* rating with logistic regression?



Estimated coefficient (logistic model): -0.82150

Multivariate logistic regression

As with linear regression, logistic regression extends to p covariates:

$$\mathrm{E}[y_i|x_i] \;=\; \frac{1}{1+e^{-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}}}$$

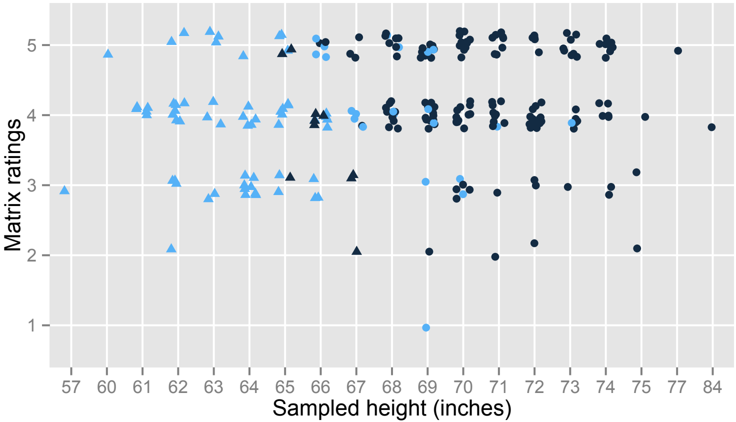
The covariates enter the probability of response through a linear combination with coefficients.

That linear combination is then passed through function μ to be appropriate as a parameter for the distribution of the response.

Does the distribution of the predictors x matter?

Classification: example

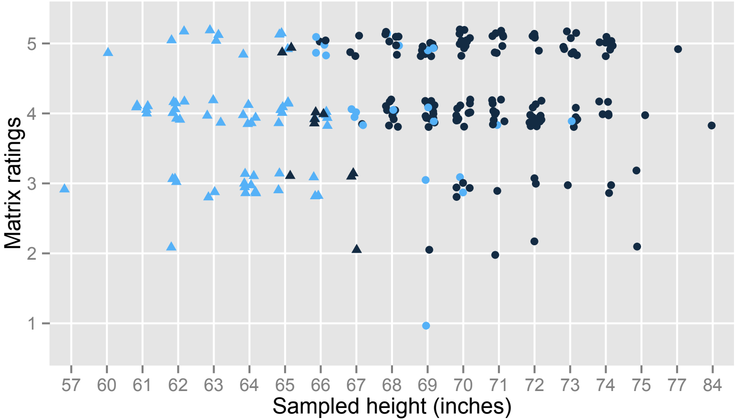
Can we classify self-reported gender using height and *Matrix* rating with logistic regression?



Estimated coefficients (logistic model): -1.0641 (height), -0.9516 (Matrix ratings)

From logistic regression to classifier

Suppose there are two covariates and two classes.



What kind of classification boundary does logistic regression create?

From logistic regression to classifier

Let's look at this equation from the point of view of classification:

$$p(y_i = 1|x_i) \quad = \quad \frac{1}{1 + e^{-\beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}}}$$

- When does $p(y = 1 \mid x) = 1/2$?
- Where does $\beta^\top x = 0$?
- What happens when $\beta^\top x < 0$?
- What happens when $\beta^\top x > 0$?

From logistic regression to classifier

Let's look at this equation from the point of view of classification:

$$p(y_i = 1|x_i) \quad = \quad \frac{1}{1 + e^{-\beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}}}$$

- When does $p(y = 1 \mid x) = 1/2$? A: When $\eta(x) = 0$
- Where does $\beta^\top x = 0$? A: A line in covariate space.
- What happens when $\beta^\top x < 0$? A: $p(y = 1 \mid x) < 1/2$
- What happens when $\beta^\top x > 0$? A: $p(y = 1 \mid x) > 1/2$

Linear separators and the margin

The classification boundary at $p(y = 1 \mid x) = 1/2$ occurs where $\beta^\top x = 0$.

Thus, logistic regression finds a *linear separator* in feature space.

Recall that SVMs also find a linear separator, but are not model based.

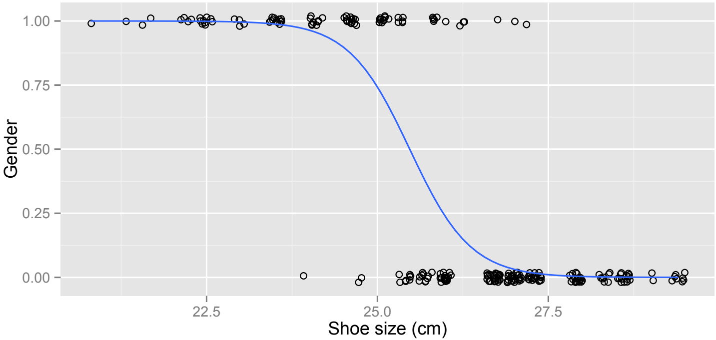
Intuitively, SVMs do not care about points that are easy to classify; rather, they try to separate the points that are difficult.

Linear separators and the margin

Loosely, logistic regression also focuses on points near the "margin"

- $\beta^\top x_i$ is the distance to the linear class separator, scaled by $||\beta||$.
- What happens to likelihood of a point further from the boundary?
- What is difference in classification between two samples one cm apart near the boundary vs far from the boundary?
- When we maximize likelihood, which points should we focus on?

Linear separators and the margin



- The probability of the class label changes most near the separator.
- When we maximize likelihood, what should we focus on? The points near the separator. They will be more informative.

Fitting logistic regression models

Data for supervised classification are $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ pairs.

Just as we did for linear regression, we fit logistic regression by maximizing the *conditional log likelihood*,

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \log p(y_i \mid x_i, \beta).$$

Objective function for logistic regression

$$\mathcal{L} = \sum_{i=1}^{n} y_i \log \mu(\beta^\top x_i) + (1 - y_i) \log(1 - \mu(\beta^\top x_i)).$$

Fitting logistic regression models

We find the derivative with the chain rule,

$$\frac{d\mathcal{L}_i}{d\beta_j} = \sum_{i=1}^{n} \frac{d\mathcal{L}}{d\mu(\beta^\top x_i)} \frac{d\mu(\beta^\top x_i)}{d\beta_j}.$$

where $\mathcal{L}_i$ is the log likelihood of the ith data point.

Objective function for logistic regression

$$\mathcal{L} = \sum_{i=1}^{n} y_i \log \mu(\beta^\top x_i) + (1 - y_i) \log(1 - \mu(\beta^\top x_i)).$$

The first term is

$$\frac{d\mathcal{L}}{d\mu(\beta^\top x_i)} = \frac{y_i}{\mu(\beta^\top x_i)} - \frac{(1 - y_i)}{1 - \mu(\beta^\top x_i)}$$

Fitting logistic regression models

Objective function for logistic regression

$$\mathcal{L} = \sum_{i=1}^{n} y_i \log \mu(\beta^\top x_i) + (1 - y_i) \log(1 - \mu(\beta^\top x_i)).$$

We use the chain rule again to compute the second term. Recall the logistic function is

$$\mu(x) = \frac{1}{1 + e^{-\eta(x)}}.$$

The derivative of the logistic with respect to its argument η is

$$\frac{d\mu(\eta)}{d\eta} = \mu(\eta)(1 - \mu(\eta)).$$

Fitting logistic regression models

Objective function for logistic regression

$$\mathcal{L} = \sum_{i=1}^{n} y_i \log \mu(\beta^\top x_i) + (1 - y_i) \log(1 - \mu(\beta^\top x_i)).$$

Now writing out the second term in the chain rule:

$$\begin{aligned} \frac{d\mu(\beta^\top x_i)}{d\beta_j} &= \frac{d\mu(\beta^\top x_i)}{d\beta^\top x_i} \frac{d\beta^\top x_i}{d\beta_j} \\ &= \mu(\beta^\top x_i)(1 - \mu(\beta^\top x_i))x_{ij} \end{aligned}$$

Conditional likelihood

Call $\mu_i = \mu(\beta^\top x_i)$. (Don't lose sight that μ_n depends on the parameter β.) The full derivative of the conditional likelihood is

$$
\begin{aligned}
\frac{d\mathcal{L}}{d\beta_j} &= \left(\frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i}\right)\mu_i(1-\mu_i)x_{ij} \\
&= (y_i(1-\mu_i) - (1-y_i)\mu_i)x_{ij} \\
&= (y_i - y_i\mu_i - \mu_i + y_i\mu_i)x_{ij} \\
&= (y_i - \mu_i)x_{ij}
\end{aligned}
$$

Gradient descent

So, the final derivative is

$$\frac{d\mathcal{L}}{d\beta_j} = \sum_{i=1}^{n}(y_i - \mu(\beta^\top x_i))x_{ij}$$

Logistic regression algorithms fit the objective with gradient methods, such as Newton's method.

Nice closed-form solutions, like the normal equations, are not available.

But the likelihood is convex; there is a unique solution.

Which samples have a lot of influence on β estimates? Which samples have little influence?

Gradient descent updates

Note that $\mathrm{E}[y_i \mid x_i] = p(y_i = 1 \mid x_i) = \mu(\beta^\top x_i)$. Recall the *linear regression* derivative. (Here, y_i is real valued.)

$$\sum_{i=1}^{n} (y_i - \beta^\top x_i) x_{ij}$$

And further recall that in linear regression, $\mathrm{E}[y_i \mid x_i] = \beta^\top x_i$. Observe that both derivatives have the form

$$\sum_{i=1}^{n} (y_i - \mathrm{E}[y \mid x_i, \beta]) x_{ij}$$

Regularized logistic regression

- We can regularize logistic regression in the same way that we regularize linear regression.
- ℓ_1-regularized logistic regression—lasso logistic regression—is used in many technologies, e.g., probably your spam filter.
- It's an efficient way to find a sparse solution.
- From the sparse solution, filtering is efficient because we need not keep track of many features.

Summary: logistic regression

- Logistic regression can be used as a binary classification method
- Logistic regression assumes linear separability and additive effects among predictors
- Inference is performed using gradient methods – closed form solution not available
- Are Bernoulli and Gaussian response distributions in regression models the only ones possible?

Additional Resources

- MLAPA Sections 8-8.3.2

- Metacademy: *Logistic regression*
- Tom Mitchell (CMU): Lecture, "Logistic regression"
- Logistic regression is a type of generalized linear model (GLM)