Non-Gaussian Regression

Justin Silverman

Penn State University

Table of Contents

Section 1

Generalized Linear Models

Generalized Linear Models

This is a very deep topic and we are just going to touch the surface as a segue to logistic regression.

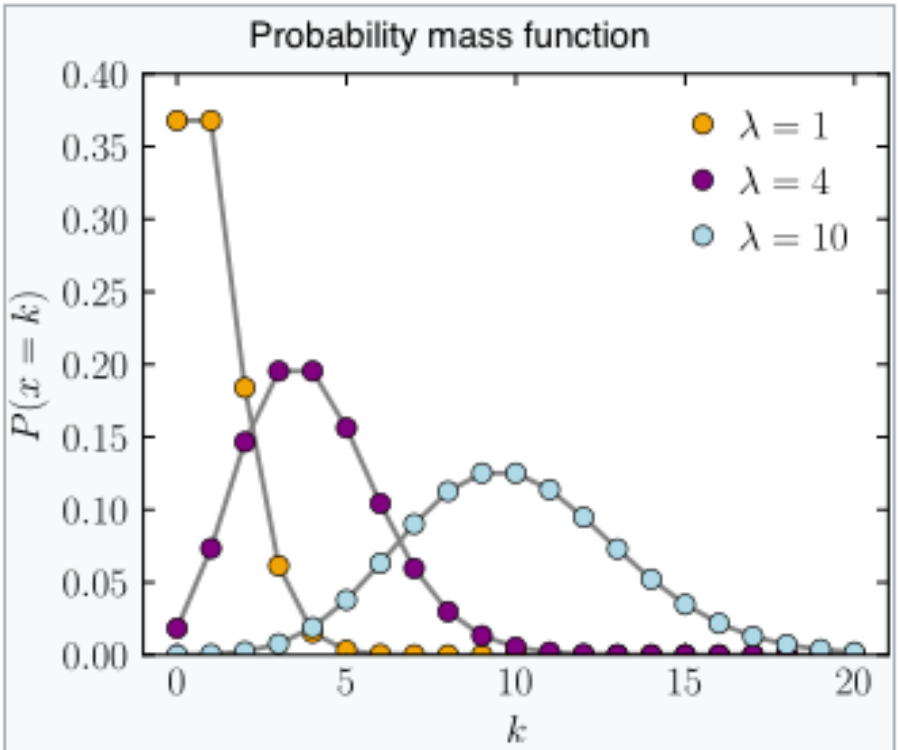
Linear Regression Revisited

$$y_i \sim N(\beta x_i, \sigma^2)$$

But what if we chose a different distribution instead of normal distribution?

The Poisson Distribution

Poisson Distribution



The horizontal axis is the index k, the number of occurrences. λ is the expected rate of occurrences. The vertical axis is the probability of k occurrences given λ. The function is defined only at integer values of k; the connecting lines are only guides for the eye.

Notation	$\mathrm{Pois}(\lambda)$
Parameters	$\lambda \in (0, \infty)$ (rate)
Support	$k \in \mathbb{N}_0$ (Natural numbers starting from 0)
PMF	$\frac{\lambda^k e^{\lambda}}{k!}$
CDF	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$, or $Q(\lfloor k+1 \rfloor, \lambda)$ (for $k \geq 0$, where $\Gamma(x, y)$ is the upper incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function)
Mean	λ
Median	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
Mode	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
Variance	λ

(From Wikipedia)

Poisson Regression (a poor choice)

So what if we did this:

$$y_i \sim \text{Poisson}(\lambda)$$
$$\lambda = \beta x_i$$

But βx_i can be negative whereas for the Poisson $\lambda \geq 0$.

Poisson Regression (a better choice)

$$y_i \sim \text{Poisson}(\lambda)$$
$$\lambda = e^{\beta x_i}$$

Note also here that $E[y_i|x_i] = e^{\beta x_i}$. We call $g(x) = e^x$ a link function in this context.

The Exponential Family

Certain probabilty distributions are part of a special family which we term the **exponential family**. These are distributions of x with parameters θ that can be written in the form

$$p(x|\theta) = h(x) \exp\left[\eta(\theta) \cdot T(x) - A(\theta)\right]$$

All you need to know is that many common distributions are in this family including:

Exponential Family Members

- Multivariate Normal, $\mathcal{R}^p$
- Exponential, $\mathcal{R}^+$
- Chi-squared, $\mathcal{R}^+$
- Gamma and Inverse Gamma, $\mathcal{R}^+$
- Beta, $\mathcal{S}^2$
- Dirichlet, $\mathcal{S}^D$
- Bernoulli, $\{0, 1\}$
- Categorical, $\{0, 1\}^K$
- Poisson, $\mathcal{Z}^+$
- Binomial(*), $\mathcal{Z}^+$
- Geometric, $\mathcal{Z}^+$
- Negative Binomial(*), $\mathcal{Z}^+$
- Multinomial(*), $\mathcal{Z}^K$
- Wishart and Inverse Wishart, (distributions over covariance matrices)

(*) only when certain parameters are fixed and known

Genearlized Linear Models typically have 3 parts

- An exponential family distribution $y \sim p(\mu, \dots)$
- A link function g such that $E(y_i|x_i) = \mu = g^{-1}(\eta)$
- A Linear predictor $\eta = \beta x$

There is really a ton of flexibility here. This is incredibly powerful:
https://en.wikipedia.org/wiki/Generalized_linear_model

Logistic Regression

We have a special name for the case where the exponential family distribution is Bernoulli and the link function is the logit function. (Logistic Regression).

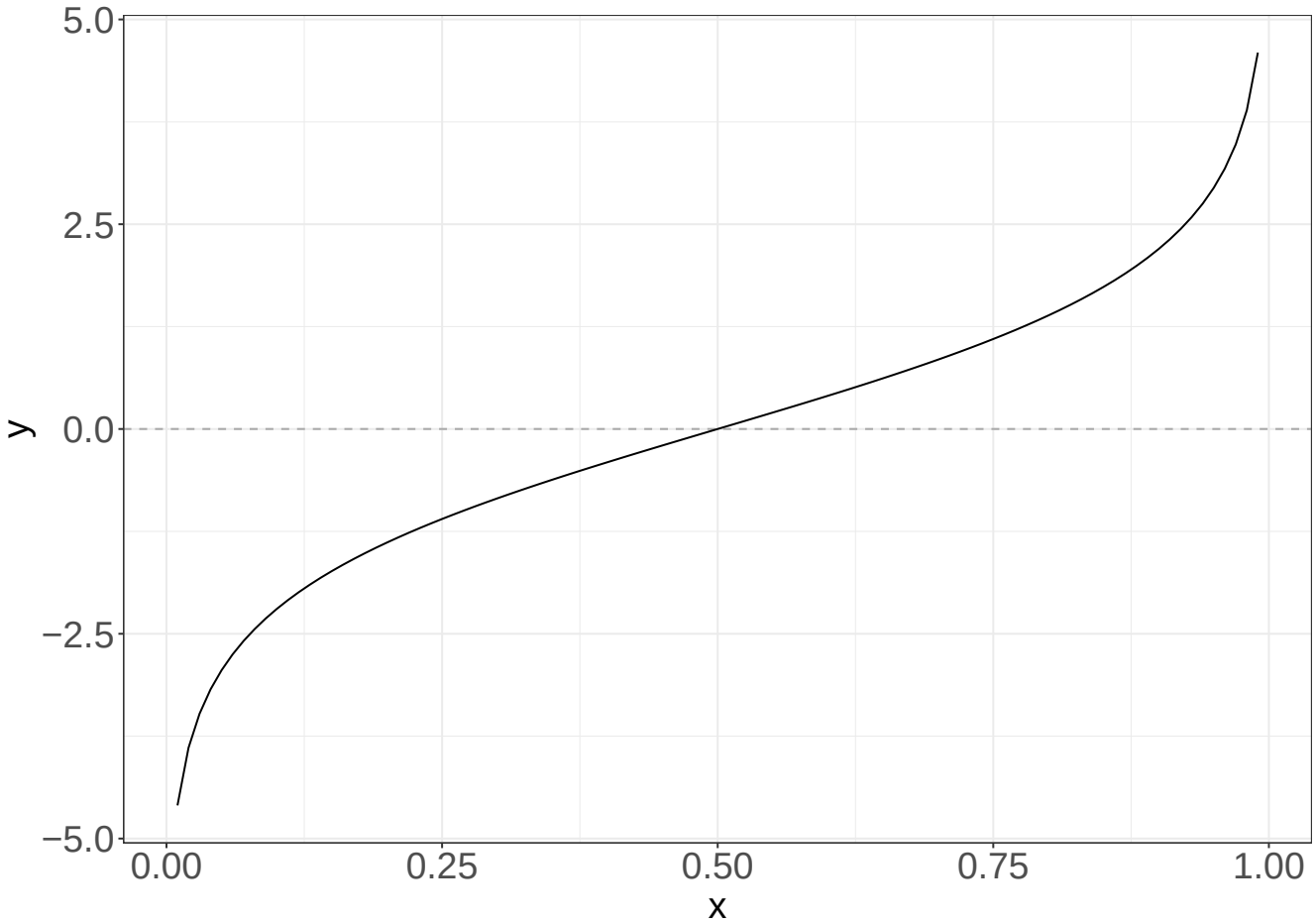
Here $y_i \in \{0, 1\}$

$$y_i \sim \text{Bernoulli}(\mu_i)$$
$$\mu_i = \text{Logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$
$$\eta_i = \beta x_i$$

That is $Logit(\mu_i) = \log \frac{\mu_i}{1-\mu_i}$. Notice that the logit function transform from $\mathcal{S}^2$ (e.g., the $(0, 1)$ interval) to $\mathcal{R}^1$.

The Logit Function

$$y = Logit(x) = \log \frac{x}{1-x}$$



Section 2

Logistic Regression for Classification

Motivation (Binary Classification)

Given appropriate covariates,

- Is that email spam or not-spam?
- Is that tumor malignant or benign?

In these cases we can encode one of the classes as 0 and the other as 1.

e.g., let $y_i = 1$ if the tumor sample i is malignant and $y_i = 0$ if its benign.

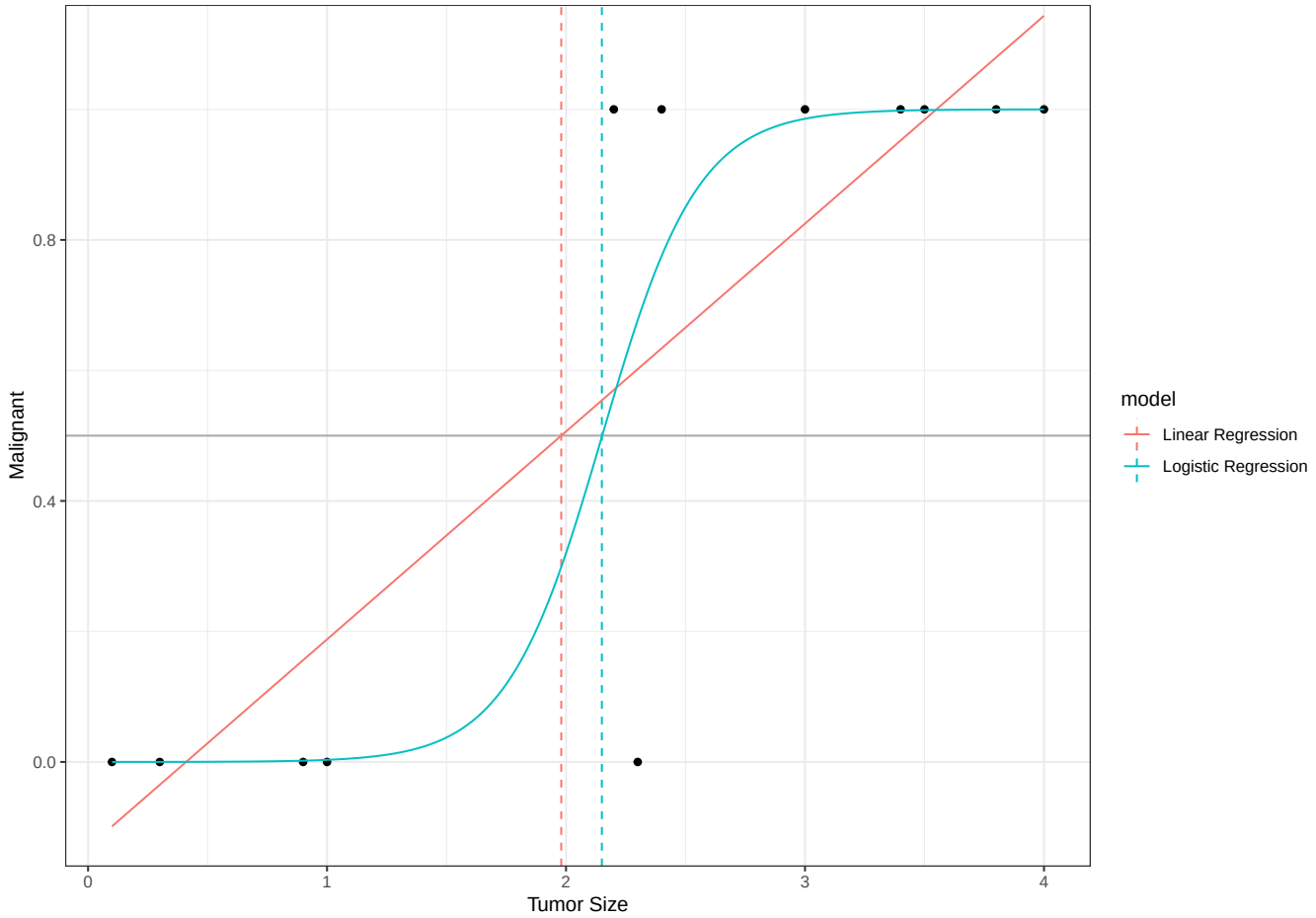
Motivation: Tumor Classification by Tumor Size

$$\text{Malignant}_i \sim \text{Bernoulli}(\mu_i)$$

$$\mu_i = \text{Logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

$$\eta_i = \beta_0 + \beta_1 \text{Tumor Size}_i$$

Linear **decision boundaries** denoted by dashed lines.

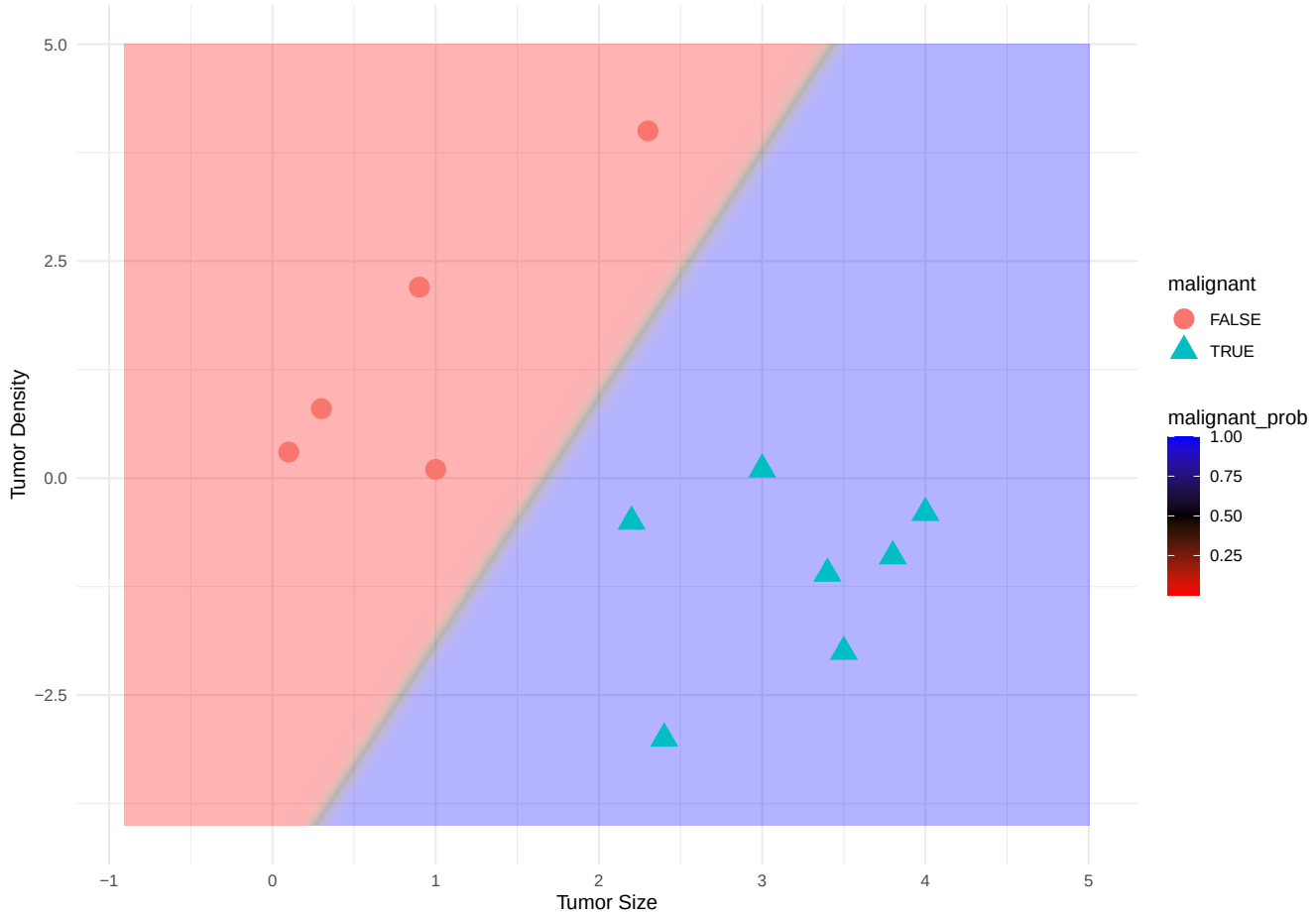


Motivation: Tumor Size and Tumor Density

$$\text{Malignant}_i \sim \text{Bernoulli}(\mu_i)$$

$$\mu_i = \text{Logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

$$\eta_i = \beta_0 + \beta_1 \text{Tumor Size}_i + \beta_2 \text{Tumor Density}_i$$

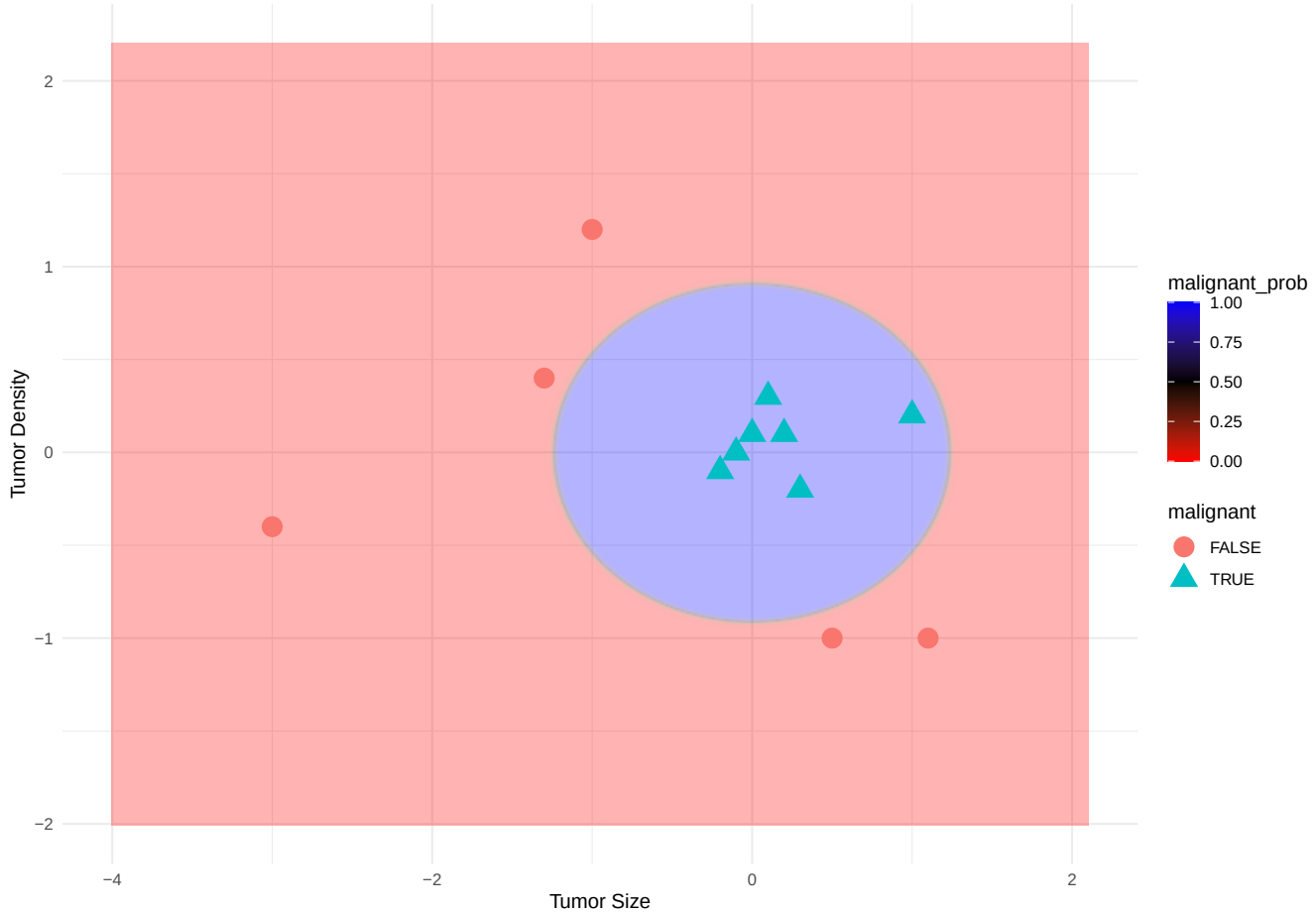


Aside: Non-linear Decision Boundaries with a Linear Model

$$\text{Malignant}_i \sim \text{Bernoulli}(\mu_i)$$

$$\mu_i = \text{Logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

$$\eta_i = \beta_0 + \beta_1 \text{Tumor Size}_i^2 + \beta_2 \text{Tumor Density}_i^2$$



Logistic Regression Loss Representation

$$y_i \sim \text{Bernoulli}(\mu_i) \to p(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i}$$
$$\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{-\eta_i}}$$
$$\eta_i = \beta x_i$$

The negative log-likelihood of this model can be written as:

$$-\log \mathcal{L}(\beta; y, x) = -\sum_{i=1}^{n} [y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]$$

where $\mu_i = 1/(1+e^{-\beta x})$.

Logistic Regression Loss Representation

However rather than coding $y_i \in \{0, 1\}$ we can instead code $\tilde{y}_i \in \{1, -1\}$ such that now $p(\tilde{y}_i = 1) = 1/(1 + e^{-\beta x})$ and $p(\tilde{y}_i = -1) = 1/(1 + e^{\beta x})$. This leads to a nicer form to work with:

$$-\log \mathcal{L}(\beta; \tilde{y}, x) = \sum_{i=1}^{n} \log \left(1 + e^{-\tilde{y}_i \beta x_i}\right).$$

In other words, the function $L(\beta) = \sum_{i=1}^{n} \log \left(1 + e^{-\tilde{y}_i \beta x_i}\right)$ is a loss function corresponding to the maximum likelihood estimate for logistic regression.

Multiclass Logistic Regression - One-vs-Rest

A simple way to extend logistic regression (really any classifier) from from the binary to the multi-class case is to **fit different classifers for each class**.

For example, if trying to classify the color a flower (red, blue, yellow) we would create three classifiers:

- red vs. [blue or yellow]
- blue vs. [red or yellow]
- yellow vs. [blue or red]

Multiclass Logistic Regression - Multinomial

aka categorical regression or softmax regression[1]

Let $y_i \in \{1, \ldots, k\}$ and x_i be a vector of p-covariates

$$
\begin{aligned}
y_i &\sim \text{Categorical}(\pi_i) \\
\pi_i &= \left(\frac{e^{\eta_{i1}}}{\sum_{j=1}^{k-1} e^{\eta_{ij}}}, \ldots, \frac{e^{\eta_{i(k-1)}}}{\sum_{j=1}^{k-1} e^{\eta_{ij}}}, \frac{1}{\sum_{j=1}^{k-1} e^{\eta_{ij}}} \right) \\
\eta_i &= \beta x
\end{aligned}
$$

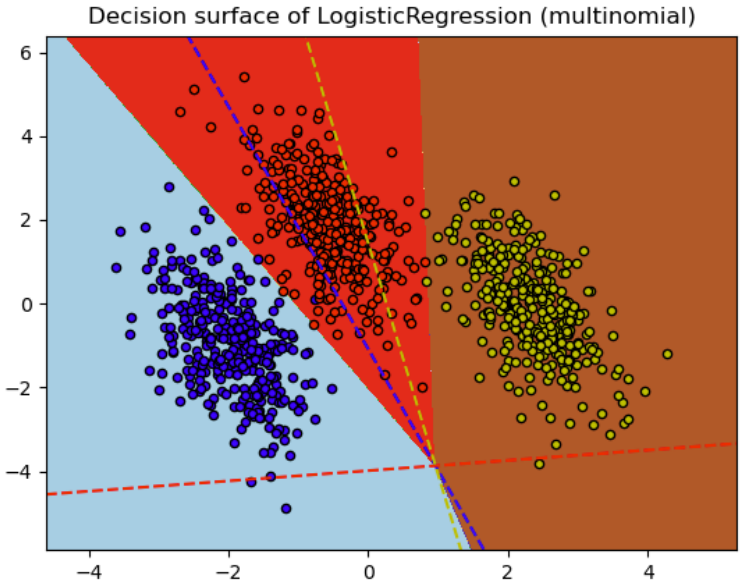
where now β is a $(k-1) \times p$ matrix.

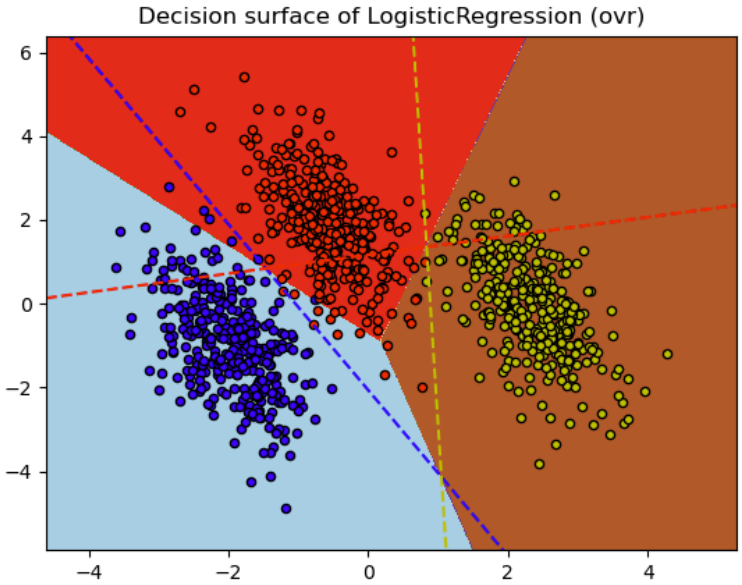
Exercise

Figure out why the numerator in the last element of π_i is 1. Hint, look back over the first lecture where compositional data was introduced.

[1] The categorical distribution is another name for the multinomial where the number of counts sums to 1, the softmax is another name for the inverse ALR or CLR transform

Multiclass Logistic Regression in Scikit-Learn





Section 3

Penalized Logistic Regression

Penalized Logistic Regression

Just as in linear regression we can create all kinds of penalized regression schemes by adding a penalty to the loss function. For example, here is a popular ℓ_2 regularization for logistic regression:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \beta x_i}\right) + \lambda ||\beta||_2$$

This can be solved using numerical optimization such as gradient descent.

Section 4

Bayesian Logistic Regresion

Bayesian Logistic Regression

We now add a prior for β

$$
\begin{aligned}
y_i &\sim \text{Bernoulli}(\mu_i) \\
\mu_i &= \text{Logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \\
\eta_i &= \beta x_i \\
\beta &\sim N(\alpha, \Sigma)
\end{aligned}
$$

Consider that the posterior distribution now represents a distribution over decision boundaries, not just a single estimated decision boundary.

In addition the prior adds, regularization / penalization like behavior as in ridge regression.

Inference: This is non-conjugate, there is no closed form solution for the posterior distribution $p(\beta|y, x)$. This can be inferred using computational methods such as MCMC or using a Laplace approximation.

Section 5

Practical Considerations

Imbalanced Classes

Numerical Optimization and Colinearity

Section 6

Questions

Space For Questions

Space for Questions