Ridge Logistic Regression for Preventing Overfitting

STA303/STA1002: Methods of Data Analysis II, Summer 2016

Michael Guerzhoy

Case Study: Images

Images are made up of pixels – tiny dots with constant colour.

A grayscale image is actually can be represented as an array of numbers between 0 and 1 (0 for black, 1 for white, numbers between 0 and 1 for different shades of gray.)

Image Classification

- Suppose we have images of 2 different people
- For a new image, want to know which of the 2 people it is
- Covariates: a vector of all the brightnesses of the grayscale image
  - X=

x0	x1	x2	x3	x4	x5	x6	x7
x12	x13	x14	x15	x16	x17	x18	x19
x24	x25	x26	x27	x28	x29	x30	x31
x36	x37	x38	x39	x40	x41	x42	x43
x48	x49	x50	x51	x52	x53	x54	x55
x60	x61	x62	x63	x64	x65	x66	x67
x72	x73	x74	x75	x76	x77	x78	x79
x84	x85	x86	x87	x88	x89	x90	x91

- Y_i: 1 if it's person A, 0 if it's person B

Logistic Regression

- $Y_i \sim Binomial(X_i \beta)$
- Find β such that the Log Likelihood is maximized
  - $\log P(y|\beta, x) = \sum_{i=1}^{m} y_i \log\left(\frac{1}{1+\exp(-x_i\beta)}\right) + (1 - y_i) \log\left(\frac{\exp(-x_i\beta)}{1+\exp(-x_i\beta)}\right)$

Classification

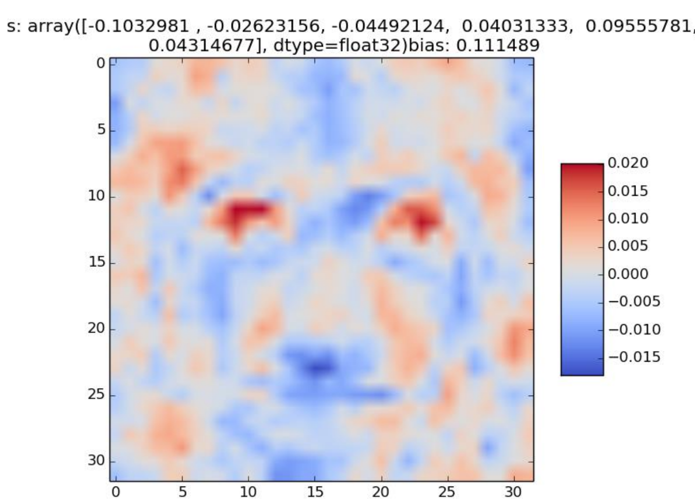
- Person A, if $X_i \beta > 0$ and Person B otherwise

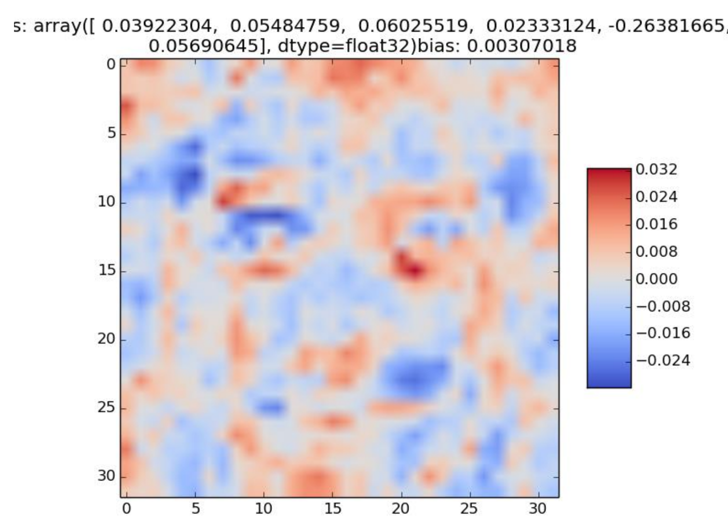
What do the βs look like?

- Remember X =

x0	x1	x2	x3	x4	x5	x6	x7
x12	x13	x14	x15	x16	x17	x18	x19
x24	x25	x26	x27	x28	x29	x30	x31
x36	x37	x38	x39	x40	x41	x42	x43
x48	x49	x50	x51	x52	x53	x54	x55
x60	x61	x62	x63	x64	x65	x66	x67
x72	x73	x74	x75	x76	x77	x78	x79
x84	x85	x86	x87	x88	x89	x90	x91

- Now go back and construct an image from the βs





Overfitting

- In a small dataset, maybe a pixel in the corner (x_1) of pictures of person A is always be smaller than 0.1, and for person B is always larger than 0.9
  - This would tend to make β_1 very large and negative
    - Recall perfect separation
  - Would hurt classification performance on new data

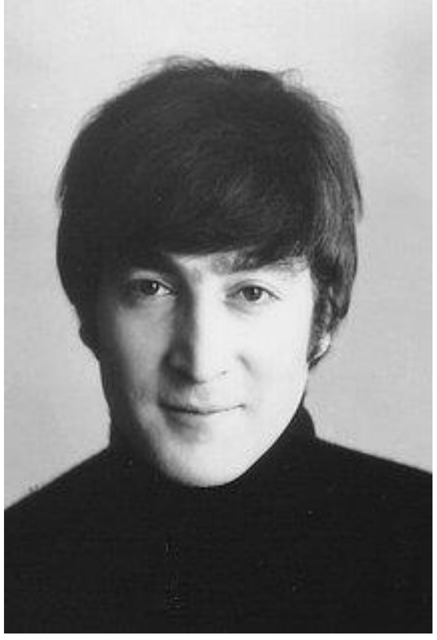
Ridge Logistic Regression

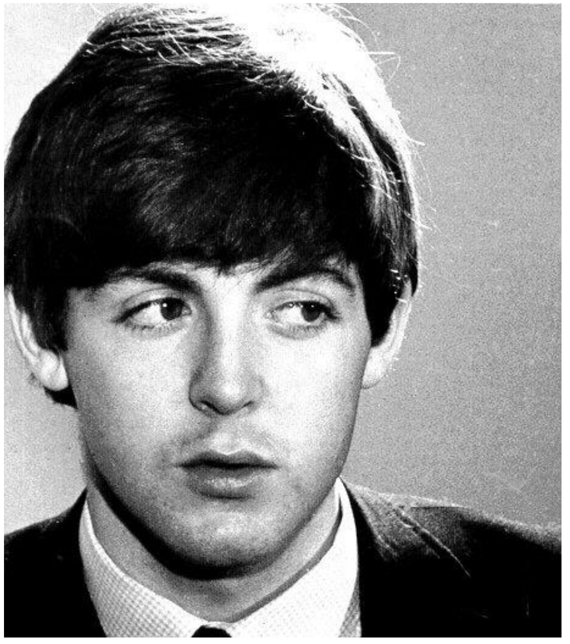
- Minimize $\mathrm{N}LL + \frac{\lambda}{2}\sum_{i=1}^{K} \beta_i^2$
  - (NLL = Negative Log-Likelihood)
- $\lambda = 0$ is what we did before
- $\lambda > 0$ means that we are *not* minimizing the NLL. Instead, we are trying to make the NLL as small as possible, while still making sure that the βs are not too large
  - Tradeoff between good fit (large log-likelihood) and good generalization (good performance on new data)
    - If we expected the "correct" βs to not be very large, makes sense to force them to be small

Ridge Logistic Regression

- Select λ using cross-validation (usually 2-fold cross-validation)
  - Fit the model using the training set data using different λ's. Use performance on the validation set as the estimate on how well you do on new data. Select the λ with the best performance on the validation set.

Ridge Logistic Regression and Inference





Is the pixel at location (20, 30) in images of John Lennon usually darker than the one in images of Paul McCartney?

- Look at the β that corresponds to the pixel at (20, 30)
- If we are using ridge regression, cannot obtain the standard error in the usual way
- If we really believe that the β cannot be too large, that *should* estimate both the standard error and the point estimate of β – people generally use the Bayesian Inference framework when using Ridge Regression