Preprocessing

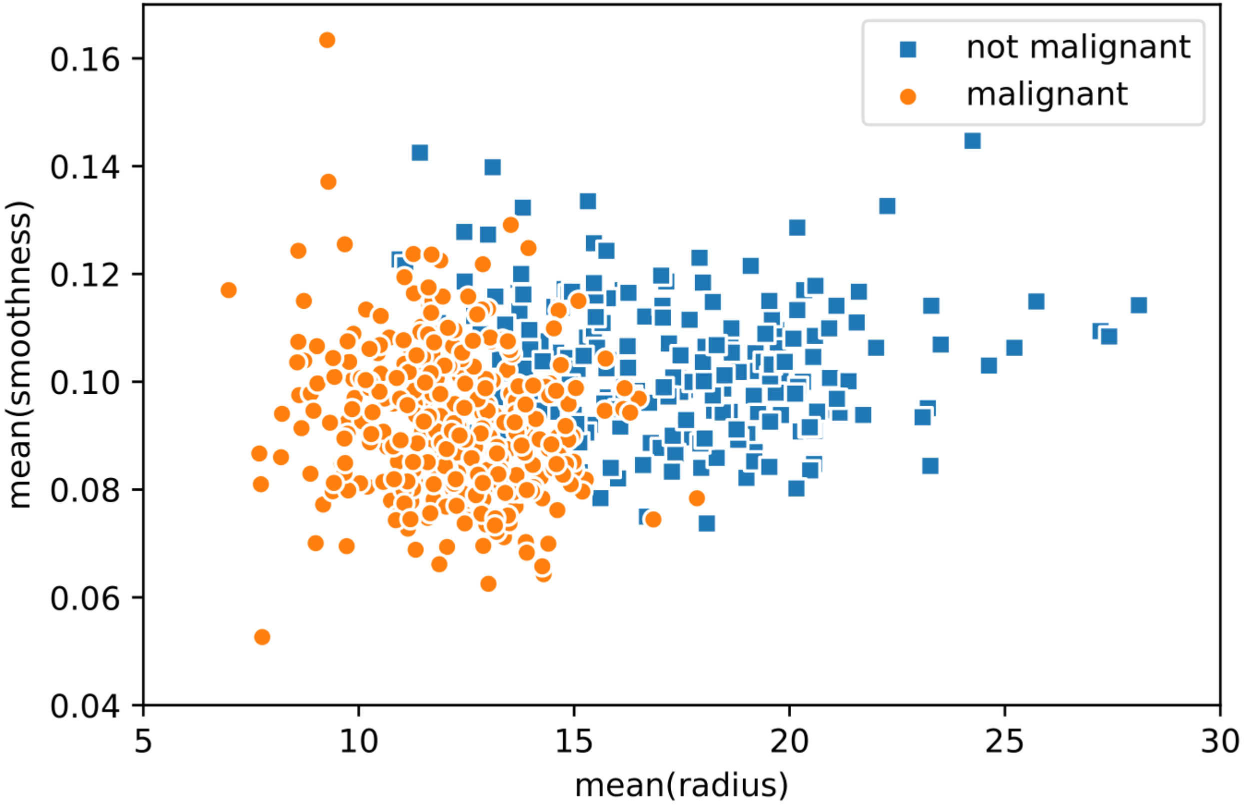
http://www.kamperh.com/

Preprocessing

Feature normalisation and scaling

http://www.kamperh.com/

Breast cancer data

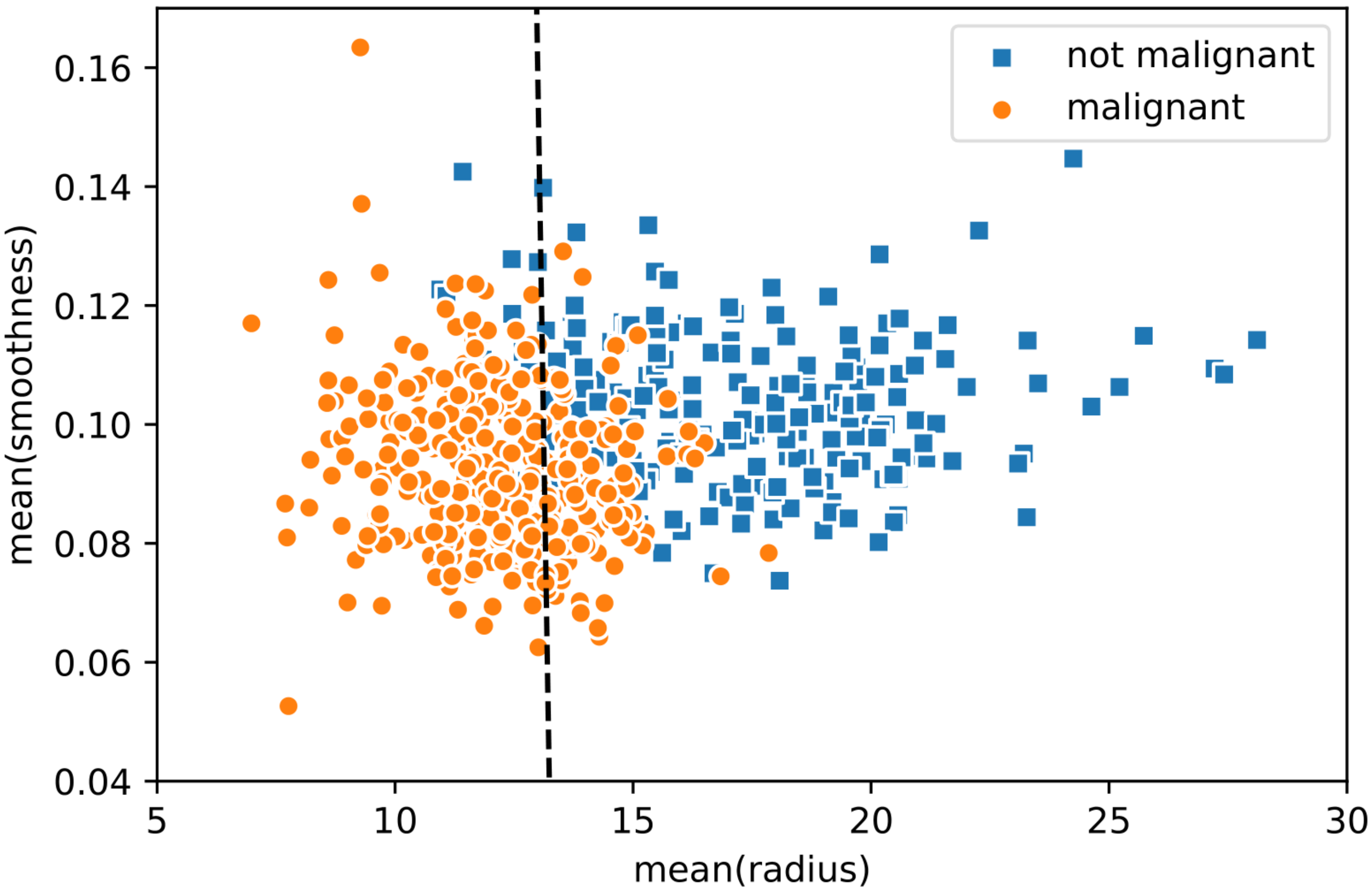


Source: `https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/`

Gradients on original data (logistic regression)

```
                          w0        w1          w2
Iteration 1000 gradients: [ -289.919 -3694.766 -26.513 ]
Iteration 2000 gradients: [ -246.909 -3223.000 -22.423 ]
Iteration 3000 gradients: [  -93.780 -1352.985  -8.034 ]
Iteration 4000 gradients: [  -92.243 -1332.636  -7.894 ]
...
```

Logistic regression on original data



Feature normalisation

Standardise the means and variances of the data:

$$\tilde{x}_d^{(n)} = \frac{x_d^{(n)} - \hat{\mu}_d}{\hat{\sigma}_d}$$

where $\hat{\mu}_d$ and $\hat{\sigma}_d^2$ are, respectively, the sample mean and variance of the d^{th} feature.

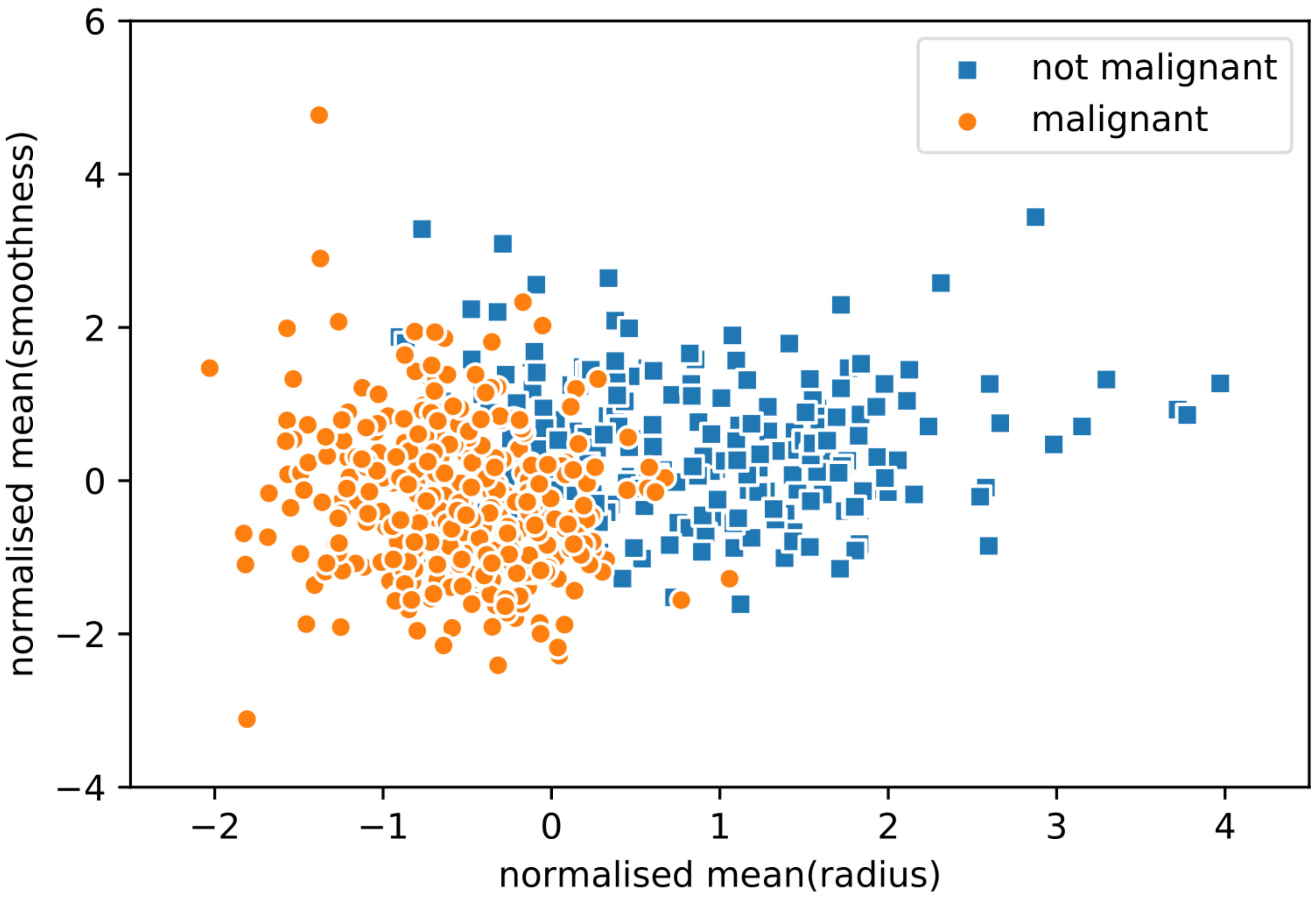
$$\hat{\mu}_d = \frac{1}{N} \sum_{n=1}^{N} x_d^{(n)}$$

$$\hat{\sigma}_d^2 = \frac{1}{N-1} \sum_{n=1}^{N} \left(x_d^{(n)} - \hat{\mu}_d \right)^2$$

$$\underset{\underline{x}^{(1)}}{\begin{bmatrix} 11 \\ 0.1 \end{bmatrix}}, \quad \underset{\underline{x}^{(2)}}{\begin{bmatrix} 25 \\ 0.12 \end{bmatrix}}, \quad \underset{\underline{x}^{(3)}}{\begin{bmatrix} 17 \\ 0.08 \end{bmatrix}}, \quad \cdots, \quad \underset{\underline{x}^{(N)}}{\begin{bmatrix} 6 \\ 0.07 \end{bmatrix}}$$

$$\underset{\tilde{\underline{x}}^{(1)}}{\begin{bmatrix} \\ \end{bmatrix}}, \quad \underset{\tilde{\underline{x}}^{(2)}}{\begin{bmatrix} \\ \end{bmatrix}}, \quad \underset{\tilde{\underline{x}}^{(3)}}{\begin{bmatrix} \\ \end{bmatrix}}, \quad \cdots, \quad \underset{\tilde{\underline{x}}^{(N)}}{\begin{bmatrix} \\ \end{bmatrix}}$$

Normalised breast cancer data

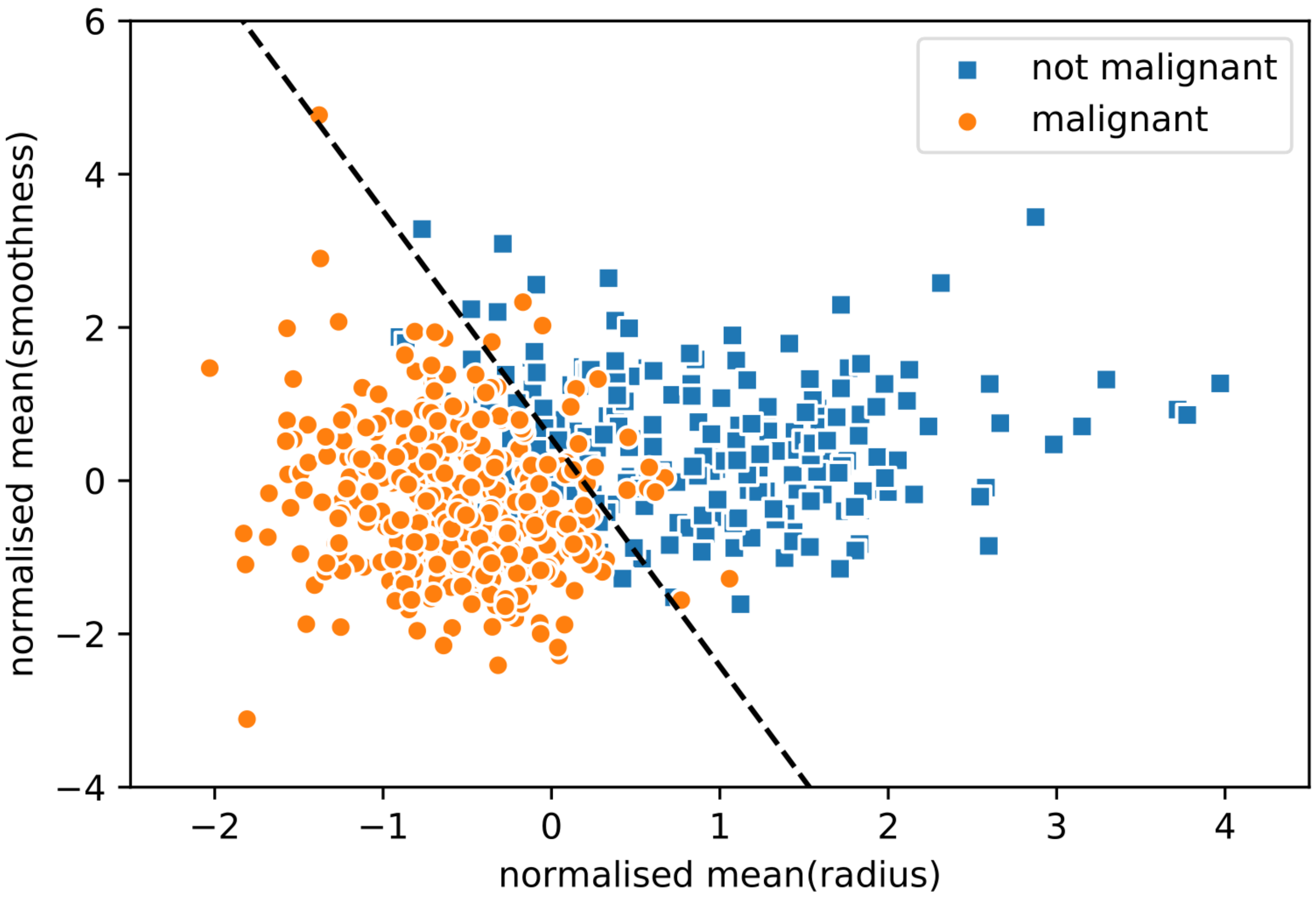


Gradients on normalised data (logistic regression)

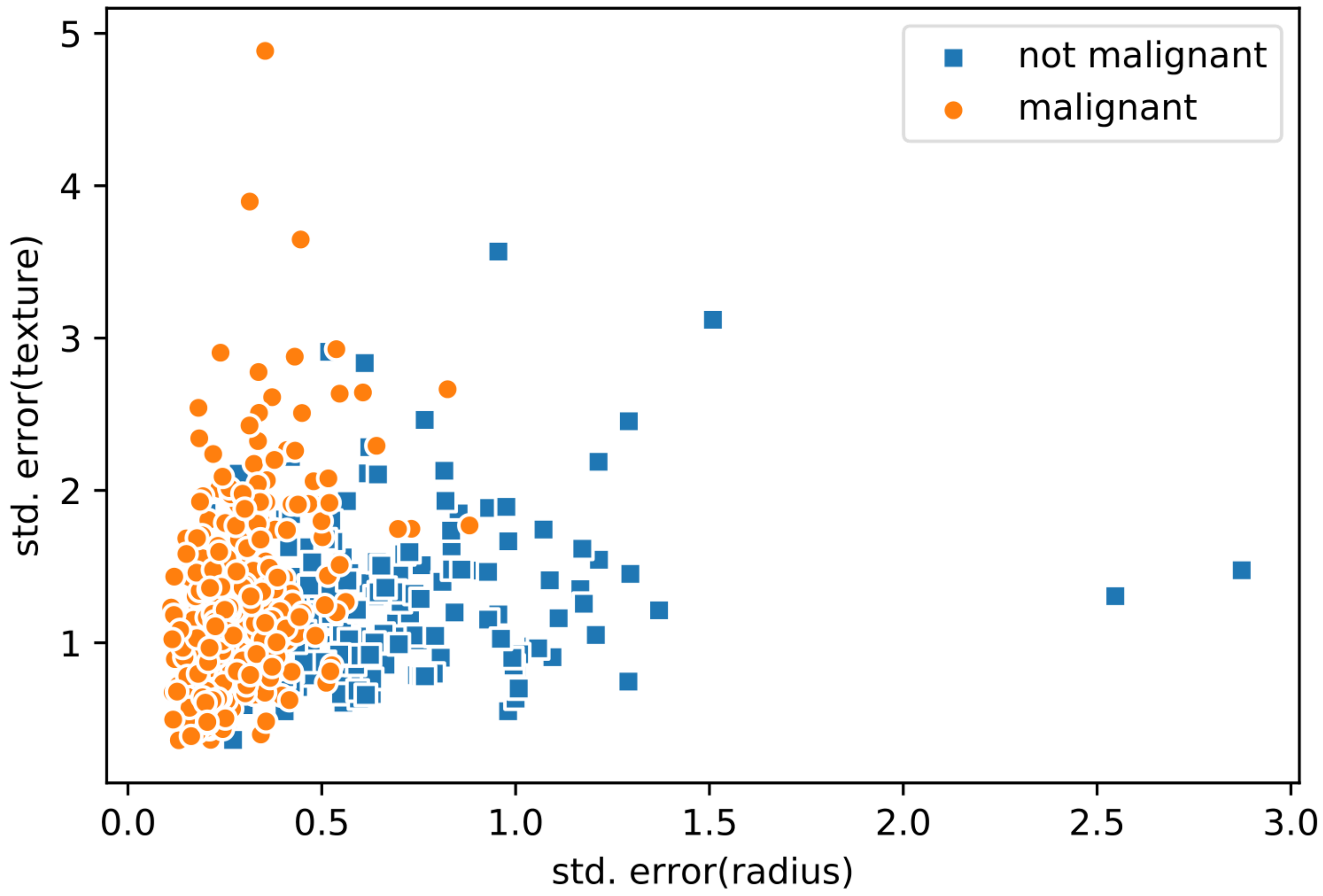
$w_0 \quad w_1 \quad w_2$

```
Iteration 1000 gradients: [ -0.525 9.179 2.472 ]
Iteration 2000 gradients: [ -0.194 3.588 0.990 ]
Iteration 3000 gradients: [ -0.096 1.752 0.486 ]
Iteration 4000 gradients: [ -0.051 0.928 0.258 ]
...
```

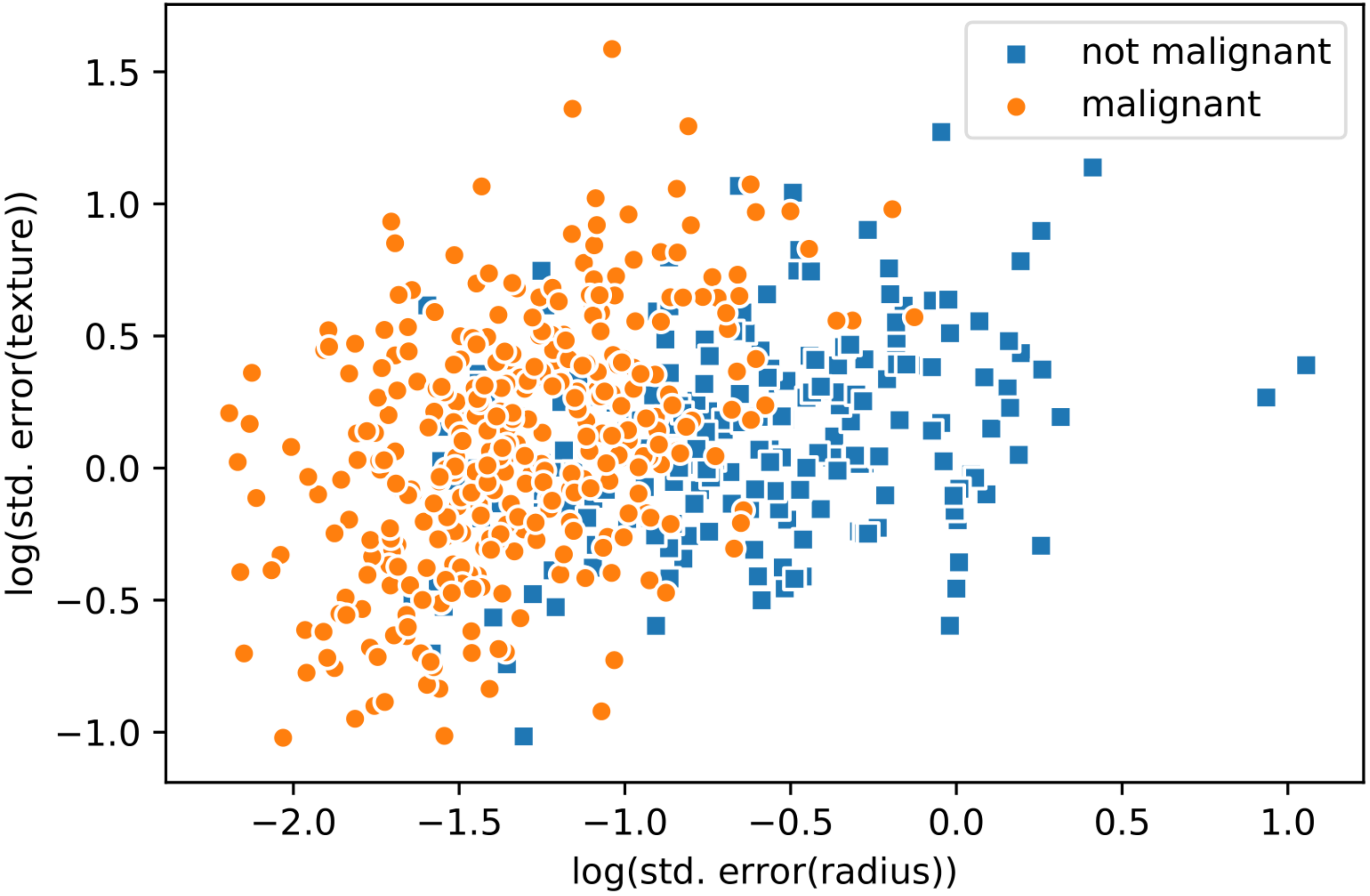
Logistic regression on normalised data



Breast cancer data



Log-scaled breast cancer data



Feature normalisation and scaling in practice

- Feature normalisation and scaling is often a bit of an art.
- You can develop an intuition as you play around with different models and optimisation algorithms.
- **Note:** Always think about how you will apply your model to new, unseen data.

Preprocessing

Categorical features and categorical output

http://www.kamperh.com/

Categorical output

- In multiclass classification we have categorical output, i.e. $y \in \{1, 2, \dots, K\}$.

- We can just save these target values explicitly. E.g. for softmax regression, you can write the loss as:

$$J(\mathbf{W}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}\{y^{(n)} = k\} \log f_k(\mathbf{x}^{(n)}; \mathbf{W})$$

- Alternatively, we can encode the target output using a *one-hot* vector:

$$\mathbf{y}^{(n)} = \begin{array}{c} \begin{array}{cccccccc} & & & & k & & & \end{array} \\ \begin{bmatrix} 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}^\top \end{array}$$

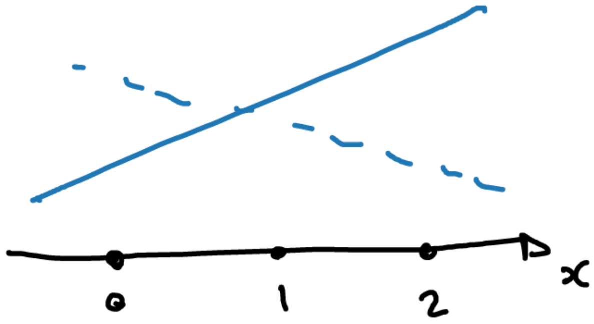
- E.g. for softmax regression, you can write the loss as:

$$J(\mathbf{W}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} y_k^{(n)} \log f_k(\mathbf{x}^{(n)}; \mathbf{W})$$

Categorical input

- Might have **inputs** that are categorical (also called *discrete* or *qualitative* features).
- E.g. someone's `occupation` might be `student`, `lecturer` or `artist`. How do we represent this?
- One option is to create a new feature:

$$x = \begin{cases} 0 & \text{if } \texttt{student} \\ 1 & \text{if } \texttt{lecturer} \\ 2 & \text{if } \texttt{artist} \end{cases}$$



- But this implies an **ordering**, which might not be true. E.g. above `artist` is closer to `lecturer` than to `student`.
- Instead use one-hot vector (also called *one-of-K*) to encode input:

$$\underset{\sim}{x} = \begin{array}{c} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \end{array} \begin{array}{l} \text{student} \\ \text{lecturer} \\ \text{artist} \end{array}$$

- Sometimes such a one-hot $\mathbf{x}$ is called a *dummy variable*.