What if you have the modeling setting where $\mathcal{Y} = \{1,2,\ldots,L\}$, a nominal categorical response with L > 2 levels. The model will still be a "classification model" but not a "binary classification model" and it's sometimes called a "multinomial classification model".

What is the null model g_0? It must only be a function of the y's

g_0 = SampleMode[y]

Consider a model that predicts on a new x_*. Let's look through D and see who's the "closest" x_i to x_* and predict yhat_* = y_i. This is called the "nearest neighbor" (NN) algorithm / model.

How do we define closest? We need a "distance function", d≥0.
For p = 1,

$$d(x_i, x_*) = \begin{cases} 0 & \text{if } x_i = x_* \\ > 0 & \text{if not} \end{cases} = |x_i - x_*| \quad L1 \text{ loss / distance}$$
$$\text{or } (x_i - x_*)^2 \quad L2$$

For p > 1,

$$d(\vec{x}_i, \vec{x}_*) = \sum_{j=1}^{p} |x_{i,j} - x_{*,j}|$$ Manhattan distance (L1)

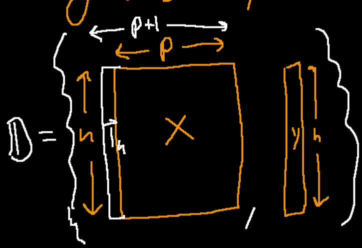
$$\text{or } \sum_{j=1}^{p} (x_{i,j} - x_{*,j})^2$$ Squared Euclidean distance (L2)
This is the default.

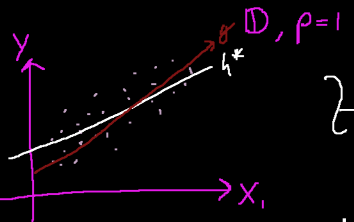
In the NN model, the function d is a hyperparameter.

An improvement on the NN model is looking at the closest K neighbors and taking the modal response of those K. This is called the KNN model where K, d are hyperparameters. default is K = √n.

Let's consider $\mathcal{Y} = \mathbb{R}$, the response is all real numbers. These models will be called "regression" which comes from historical circumstance only. What is the null model?

$g_0(\vec{x}_*) = \bar{y}$





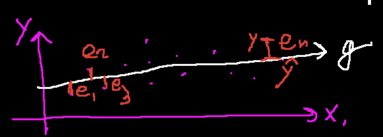
$$\mathcal{H} = \left\{ \underbrace{\vec{w} \cdot \vec{x}}_{w_0 + w_1 x_1 + \ldots + w_p x_p} : \vec{w} \in \mathbb{R}^{p+1} \right\}$$

Linear Regression since the candidate set is the set of all lines.

$$h^*(\vec{x}) = \underbrace{w_0^*}_{\beta_0} + \underbrace{w_1^*}_{\beta_1} x_1 + \ldots + \underbrace{w_p^*}_{\beta_p} x_p \Rightarrow y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

ignorance + misspecif. error

Let p = 1. How do we compute w_0 and w_1? I.e. what is $\mathcal{A}$? Algorithms usually minimize an objective / loss / error / fitness function which captures errors which are model mistakes.



$$e_i := y_i - \hat{y}_i = y_i - g(\vec{x}_i)$$

What are some reasonable objective functions?

$$SAE := \sum_{i=1}^{n} |e_i|$$ L1 loss, sum of absolute error

$$SSE := \sum_{i=1}^{n} e_i^2$$ L2 loss, sum of squared error (default)

$$\mathcal{A}: b_0, b_1 = \underset{w_0, w_1 \in \mathbb{R}}{\text{argmin}} \{SSE\}$$

This can be computed analytically (i.e. there is a mathematically derivable answer).

"Least squares regression"

$$\frac{\partial}{\partial w_0}[SSE] = \frac{\partial}{\partial w_0}\left[\sum (y_i - w_0 + w_1 x_{i,1})^2\right] = \frac{\partial}{\partial w_0}[\;] = \cancel{2} n w_0 - \cancel{2} n\bar{y} + \cancel{2} w_1 n \bar{x} \overset{set}{=} 0$$
$$\Rightarrow b_0 = \bar{y} - b_1 \bar{x}$$

$$\sum y_i^2 + w_0^2 + w_1^2 x_i^2 - 2 y_i w_0 - 2 y_i w_1 x_i + 2 w_0 w_1 x_i$$
$$= \sum y_i^2 + n w_0^2 + w_1^2 \sum x_i^2 - 2 w_0 \underbrace{\sum y_i}_{n\bar{y}} - 2 w_1 \sum x_i y_i + 2 w_0 w_1 \underbrace{\sum x_i}_{n\bar{x}}$$

$$\frac{\partial}{\partial w_1}[SSE] = \frac{\partial}{\partial w_1}[\;] = \cancel{2} w_1 \sum x_i^2 - \cancel{2} \sum x_i y_i + \cancel{2} w_0 n\bar{x} \overset{set}{=} 0$$

$$\Rightarrow \sum x_i^2 w_1 = \sum x_i y_i - w_0 n\bar{x} \Rightarrow w_1 \sum x_i^2 = \sum x_i y_i - \underbrace{(\bar{y} - w_1 \bar{x}) n\bar{x}}_{-n\bar{x}\bar{y} + n\bar{x}^2 w_1}$$

$$\Rightarrow w_1 \left(\sum x_i^2 - n\bar{x}^2\right) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$\Rightarrow b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\frac{1}{n-1}\left(\sum x_i y_i - n\bar{x}\bar{y}\right)}{\frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right)} = \frac{r s_x s_y}{s_x^2} = r\frac{s_y}{s_x} \Rightarrow b_0 = \bar{y} - r\frac{s_y}{s_x}\bar{x}$$

FYI from a stat-101 class...

$$\rho = \text{Corr}[X, Y] = \frac{\text{Cov}[X,Y]}{SD[X]\,SD[Y]} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{SD[X]\,SD[Y]}$$

Covariance is estimated with sample covariance:

$$s_{XY} := \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum x_i y_i - n\bar{x}\bar{y} - \cancel{n\bar{x}\bar{y}} + \cancel{n\bar{x}\bar{y}}\right) = \frac{1}{n-1}\left(\sum x_i y_i - n\bar{x}\bar{y}\right)$$

Correlation is estimated with sample correlation:

$$r = \frac{s_{XY}}{s_X s_Y} \Rightarrow s_{XY} = r s_X s_Y$$

Variance is estimated with sample variance:

$$s_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right)$$