3) Elastic Net

is a combination of L2 and L1.

The implementation has both λ_1 and λ_2 for L1 and L2 respectively same $\lambda_1 = \lambda_2$

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1$$

In logistic regression the $|y - X\beta|^2$ is $J(\beta)$

$$J(\beta) = -y_i \log h(\beta) - (1 - y_i) \log (1 - h(\beta))$$

β is the weight for features we need to learn.

So

$$L(\lambda_1, \lambda_2, \beta) = \hat{\sum_{i=1}} -y_i \log h(\beta) - (1 - y_i) \log (1 - h(\beta)) + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1$$

Where $\frac{L(\lambda_1, \lambda_2 \beta)}{\partial \beta}$ $\Rightarrow$ to find the maximum solution for gradient descent stochastic approach for β

$$|\beta|^2 = \sum_{k=1}^{d} \beta_k^2 \qquad |\beta|_1 = \sum_{k=1}^{d} |\beta_k|$$

d is number of features

Gradient descent step:

$$\frac{\partial L}{\partial w_{ij}} = \sum_{i=1}^{\hat{}} x_{ij}\left(y^{i} - \hat{y}^{i}(\beta)\right) + \lambda_2|\beta| + \frac{\lambda_1|\beta|}{\beta}$$

The gradient descent step is divided by samples to average out.

The L2 is applied with the data and after that L1 is applied on the previous weights without updating ~~the~~ weights learned from L2. The elastic net estimator is a 2-stage ~~es~~ procedure: for fixed λ_2 we first find ridge regression coefficients (β_{L_2}) and then we do the lasso shrinkage along the lasso coefficient~~s~~ solution path.