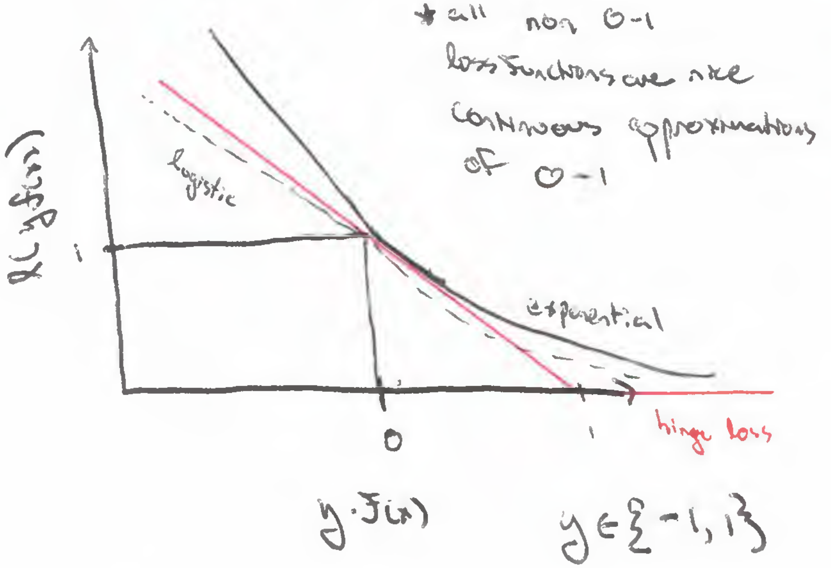
Loss Functions

- $F(x)$ is a signed "score function" and y is class label. Thus we want loss functions which penalize a lot for very negative values of $yF(x)$
- 0-1 (minimizes misclassification rates, but not differentiable at origin and has no gradient)

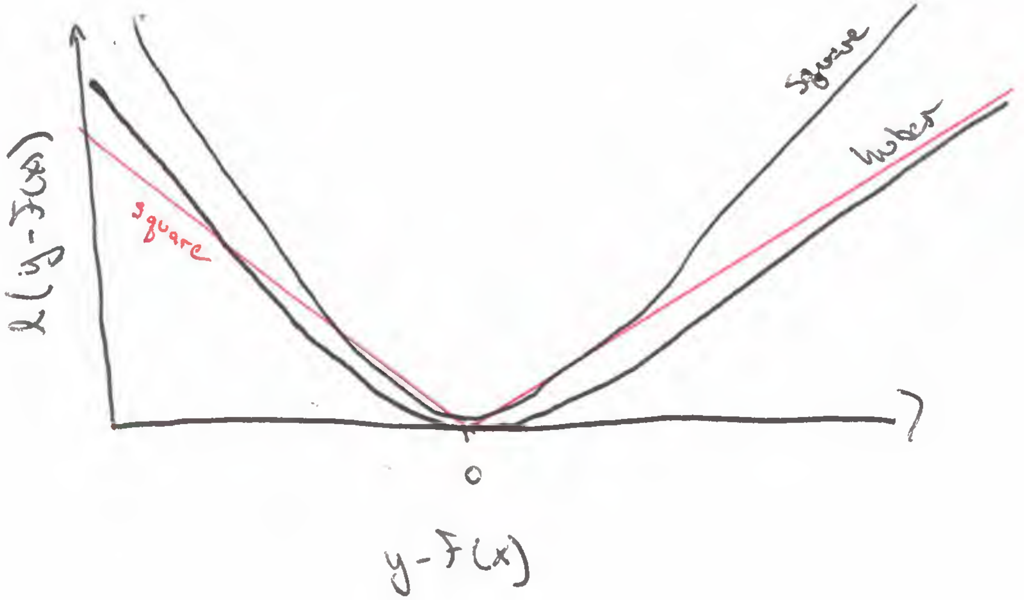
classification losses



- exponential (nice theoretical properties, but increases exponentially so is sensitive to outliers
- hinge (SVM loss)
- logistic (logistic regression loss)

- square (differentiable at origin, gradient slows down at origin which is helpful in gradient decent, but is sensitive to outliers since it increase quadratically).
- absolute (less sensitive to outliers since it increase linearly but, not differentiable at origin, and large gradient at origin).

regression losses



- huber (quadratic at origin and linear away from origin, the best of square and absolute loss, have 1 more hyperparameter, δ, to tune).

- Some nice theoretical justifications for using l_2 and l_1 loss.

  - the optimal prediction function using square loss results in predicting the conditional expected mean: $f^*(x) = \mathbb{E}[Y|X=x]$

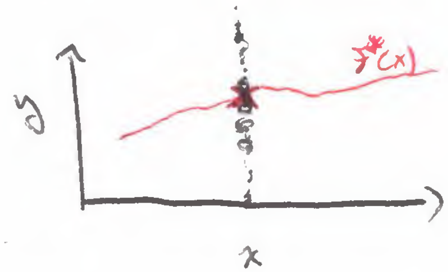
minimal expected loss for each individual x

$$0 = \frac{d}{d\hat{y}} E[(Y-\hat{y})^2|X] = \int \frac{d}{d\hat{y}}(y-\hat{y})^2 P(y|x)\,dy \Rightarrow \hat{y}^* = \int y\,p(y|x)\,dy = E[Y|X]$$

$$\Rightarrow \text{for each } x, \quad \boxed{f^*(x) = E[Y|X=x]}$$

minimal expected loss for all xs

- using data, the prediction is thus the mean conditional estimator



  - the optimal prediction function using absolute loss results in predicting the conditional median

$$0 = \frac{d}{d\hat{y}} E[|Y-\hat{y}||X] = \int_{-\infty}^{\hat{y}} \frac{d}{d\hat{y}}(\hat{y}-y) P(y|x)\,dy + \int_{\hat{y}}^{\infty} \frac{d}{d\hat{y}}(y-\hat{y}) P(y|x)\,dy$$

$$\Rightarrow \int_{-\infty}^{\hat{y}^*} P(y|x)\,dy = \int_{\hat{y}^*}^{\infty} P(y|x)\,dy$$

$\Rightarrow$ the only place this happens is at the median $\Rightarrow \hat{y}^* = \text{med}(P(y|x))$

$$\Rightarrow \boxed{f^*(x) = \text{med}(P(Y=y|X=x))}$$

- using data, the prediction is the median conditional estimator