Logistic regression:

- Nice, useful model binary classifier

Model

$$\log \frac{P(1|x)}{P(-1|x)} = w^T x$$

↑ you can have a constant here by extending x

So $$P(1|x) = \frac{e^{w^T x}}{1 + e^{w^T x}}$$

We can fit with maximum likelihood

$$\ell(w) = \sum_{i \in \text{examples}} \log P(y_i|x_i)$$

this is equiv to minimizing -ve log-likelihood

$$w = \underset{w}{\operatorname{argmin}} - \sum_{i \in \text{examples}} \log P(y_i | x_i)$$

$$= \underset{w}{\operatorname{argmin}} - \sum_i \left[\frac{(y_i + 1)}{2} w^T x_i - \log\left(1 + e^{w^T x_i}\right) \right]$$

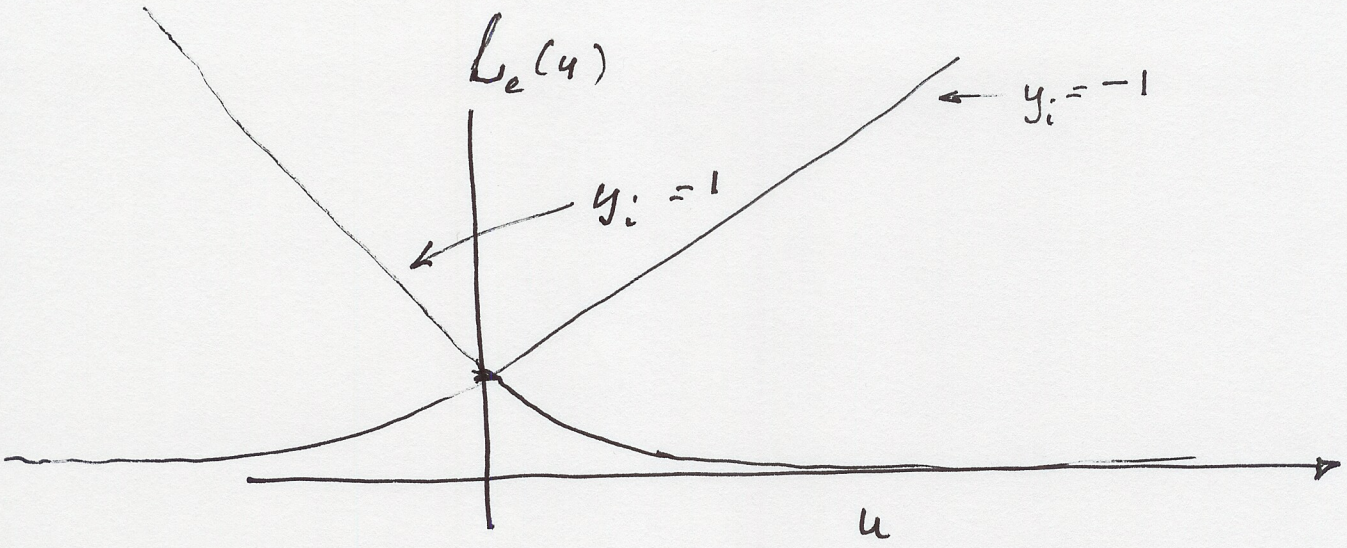
Now regard

$$\log\left(1 + e^{u_i}\right) - \frac{(y_i + 1)}{2} u_i \quad \text{as a loss}$$

$L(y_i, u_i)$ ← exponential loss

y_i: Value of example

u_i: Score $= w^T x_i$



looks a lot like hinge loss:

So we can think of LiR as minimizing a loss function.

How? $\nabla = 0$, Newton.

$$\nabla_w L_e = \sum_i \left[\frac{x_i}{1 + e^{w^T x_i}} - \frac{(y_i + 1)}{2} x_i \right]$$

$$H_w L_e = \sum_i \frac{e^{w^T x_i}}{(1 + e^{w^T x_i})^2} \cdot x_i x_i^T$$

Notice:

- examples where $w^T x_i$ has large abs value have little effect on H.
- For ~~these~~ others, H looks like a covariance
- what if features are correlated?
  - H has small eigenvalues
  - esp if w will be unstable in these dirs.
  - should manifest as large w

$\Rightarrow$ Regularize.

Solve

$$\underset{w}{\operatorname{argmin}} \quad L_e(w, \text{examples}) + \lambda \|w\|$$

↑ ?

Some norm