Back to the cost functions

Regularization term

Logistic Regression

$$J(\theta) = \frac{1}{n}\sum_{i=0}^{n} \log_2\left(1 + \exp(-y_i(\theta \cdot x_i))\right) + \frac{\lambda}{2}||\theta||^2$$

Support Vector Machines

$$J(\theta) = \frac{1}{n}\sum_{i=0}^{n} \max(1 - y_i(\theta \cdot x_i), 0) + \frac{\lambda}{2}||\theta||^2$$

Quick Intuition for Regularization

Consider high-dimensional problems

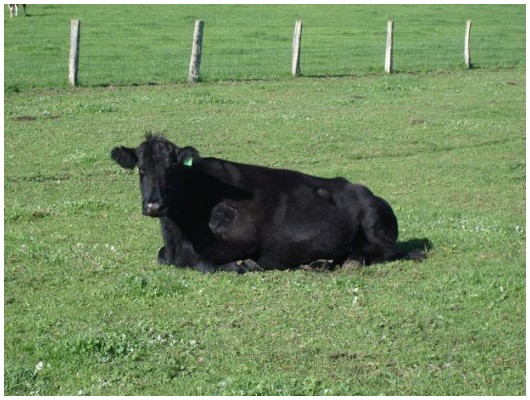
- Image recognition

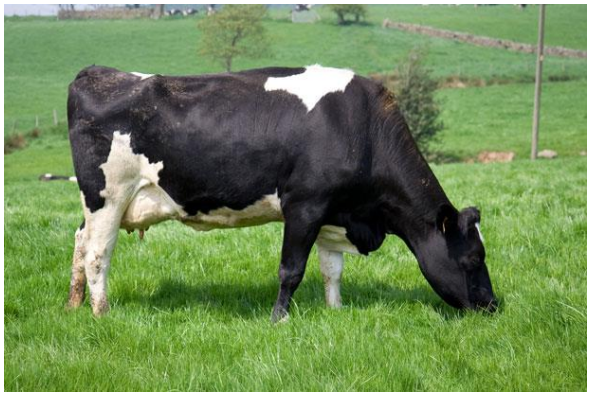
  one weight per pixel

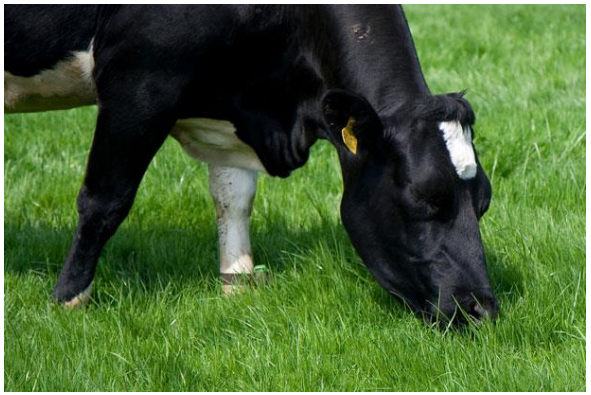
- Document vectors

  500k term weights

With so many weights, likely that many are correlated







Nearby pixels are highly correlated.

As one weight goes up, another goes down to compensate.

And so weights may explode – overfitting again

Need to enforce some condition on the weights to prefer simple models.

The regularization term provides this balance

Aside on Norms

Norm: any function that assigns a strictly positive number to every non-zero vector

$$L^p\text{-norm} = ||x||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

Aside on Cost Functions

not a norm $$\sum_i H_i - H'_i$$

errors cancel out;
usually not what you want

L^1-norm $$\sum_i |H_i - H'_i|$$

Asserts that 1 error of 7 units is as bad as 7 errors of 1 unit each

L^2-norm $$\sqrt{\sum_i (H_i - H'_i)^2}$$

Asserts that 1 error of 7 units is as bad as 49 errors of 1 unit each

$$\frac{1}{n} \sum_i^n (H_i - H'_i)^2$$

Average squared error per data point; useful when comparing methods that filter the data differently

Back to Regularization

$$||\theta||_1 = \sum_i |\theta_i|$$

"LASSO"
Regularized Least Squares; L1 norm

$$\alpha \sum_{k=1}^{n} (h(x_k) - y_k)^2 + \frac{\lambda}{2} ||\theta||_1$$

$$||\theta||_2 = \sqrt{\sum_i \theta_i^2}$$

"Ridge Regression"
Regularized Least Squares; L2 norm

$$\alpha \sum_{k=1}^{n} (h(x_k) - y_k)^2 + \frac{\lambda}{2} ||\theta||_2^2$$