CS195-5 : Introduction to Machine Learning
Lecture 13

Greg Shakhnarovich

October 4, 2006

Announcements

Review

- Batch gradient descent:

$$\mathbf{w}_{new} := \mathbf{w} + \mathbf{H}^{-1}\frac{\partial}{\partial \mathbf{w}}\ell(X_N; \mathbf{w}),$$

  where $\mathbf{H}$ is the Hessian matrix (second derivatives) of $\ell(X_N; \mathbf{w})$.

- Batch vs. stochastic GD.

- Overfitting with logistic regression on separable training sets.

Today

- Gradient descent demo (take II)
- Regularization: a method for preventing overfitting
  - MAP estimation
  - Gaussian prior $\Rightarrow L_2$ regularization;
  - Laplacian prior $\Rightarrow L_1$ regularization.

Softmax

- The multiclass generalization of the logistic model:

$$p\left(y = c \,|\, \mathbf{x}\right) \;=\; \frac{\exp\left(\mathbf{w}_c^T \mathbf{x}\right)}{\sum_{k=1}^{C} \exp\left(\mathbf{w}_k^T \mathbf{x}\right)}$$

- For $C = 2$, this is identical to the logistic regression (PS2).

- The boundaries between classes stil linear.
  - Can be extended to nonlinear by using nonlinear basis functions.

- Why is it called "the *softmax* model"?

Overfitting with logistic regression

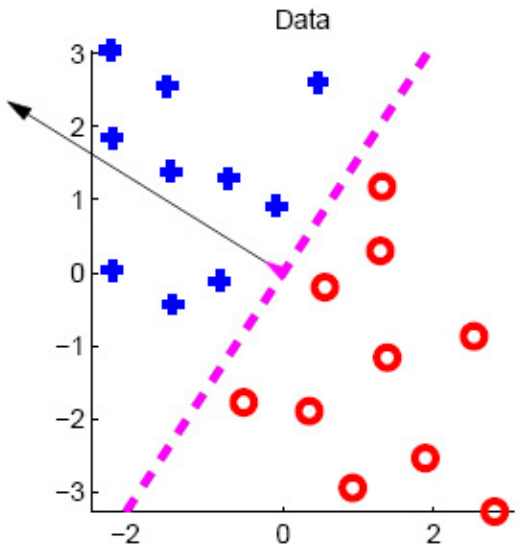
- We can get the same decision boundary with an infinite number of settings for $\mathbf{w}$.
- When the data are *separable* by $w_0 + \alpha \mathbf{w}^T \mathbf{x} = 0$, what's the best choice for α?

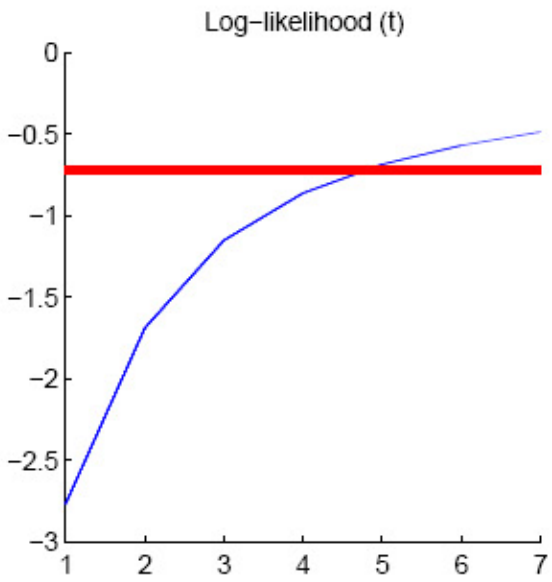
$$p(y = 1 \,|\, \mathbf{x}) \;=\; \sigma(w_0 + \alpha \mathbf{w}^T \mathbf{x}).$$

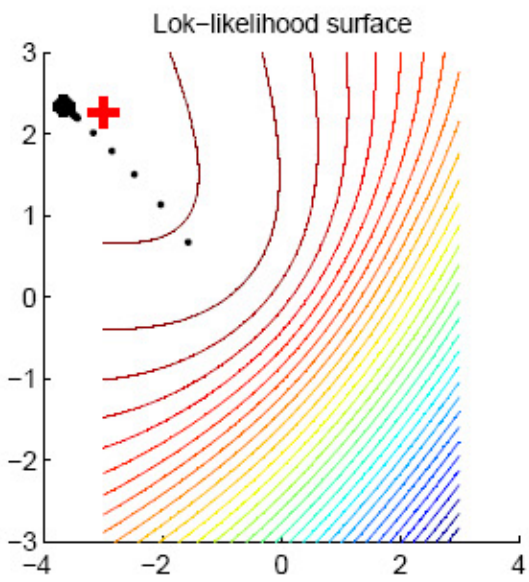
- With $\alpha \to \infty$, we have $p(y_i | \mathbf{x}; w_0, \alpha \mathbf{w}) \to 1$.
- With $\alpha = \infty$ there is a continuum of w_0 that reach perfect separation.
- When the data are not separable, similar effect is present but more subtle.

Batch GD demo, take II

Batch GD demo, take II







MAP estimation for logistic regression

- Intuition: similar to the coin toss experiment, we may have some belief about the value of $\mathbf{w}$ before seeing any data.
  - E.g., may prefer smaller values of $\|\mathbf{w}\|$.
- A possible prior that captures that belief:

$$p(\mathbf{w}) \;=\; \mathcal{N}\left(\mathbf{w};\, \mathbf{0}, \sigma^2 \mathbf{I}\right).$$

- In the 2D case (again, ignoring w_0) this means

$$p(w_1, w_2) \;=\; \frac{1}{2\pi\sigma^2} \exp\left(-\frac{w_1^2 + w_2^2}{2\sigma^2}\right).$$

MAP for logistic regression

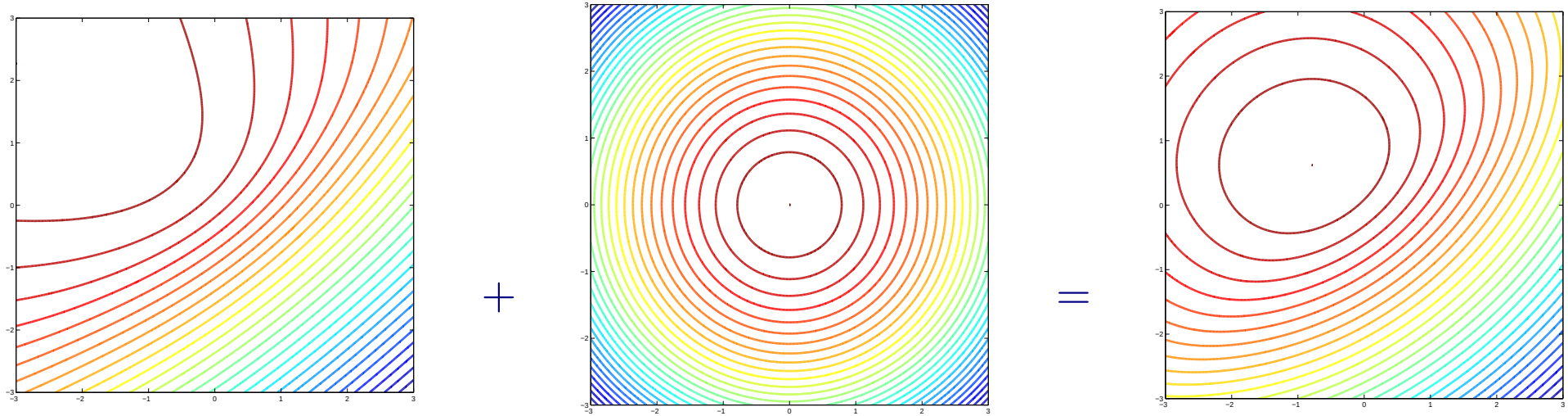
- Instead of $\ell(X_N; \mathbf{w})$ the objective function (under the Gaussian prior) becomes:

$$\begin{aligned}\tilde{\ell}(X_N; \mathbf{w}, \sigma) &= \ell(X_N; \mathbf{w}) + \log p(\mathbf{w}) \\ &= \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}\right) - \frac{1}{2\sigma^2}(w_1^2 + w_2^2) + \text{const}(\mathbf{w}).\end{aligned}$$

- This is a *penalized log-likelihood* (or *log-posterior*).
- Note that $w_1^2 + w_2^2 = \|\mathbf{w}\|^2$.
- Setting σ^2 will affect the penalty we impose for a particular value of $\|\mathbf{w}\|$.

Penalized likelihood surface

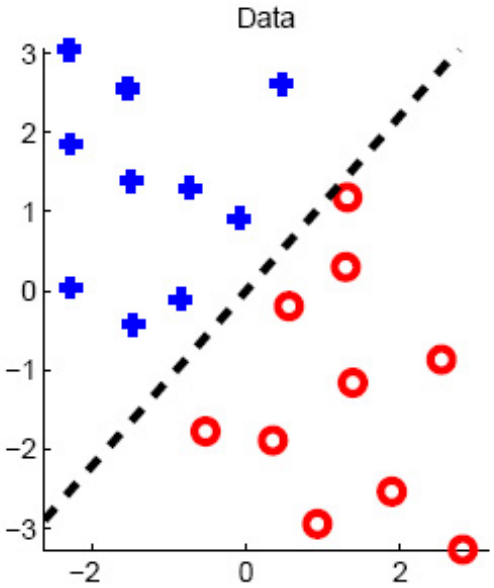
$\ell(X_N; \mathbf{w})$ $\log p(\mathbf{w}; \sigma)$ $\tilde{\ell}(X_N; \mathbf{w}, \sigma)$

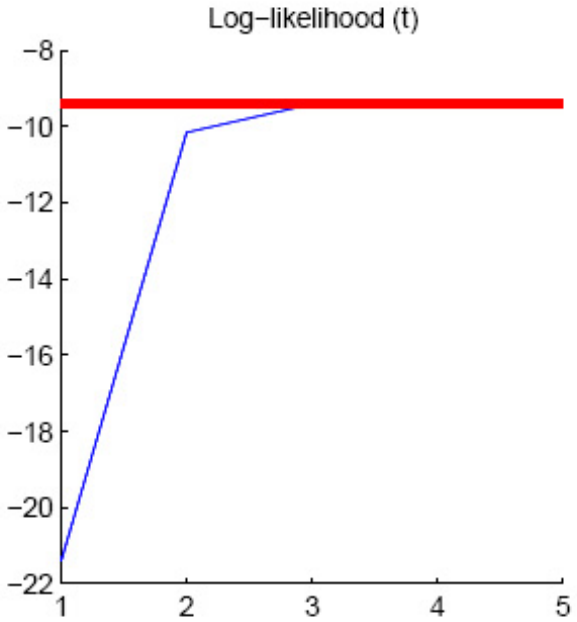


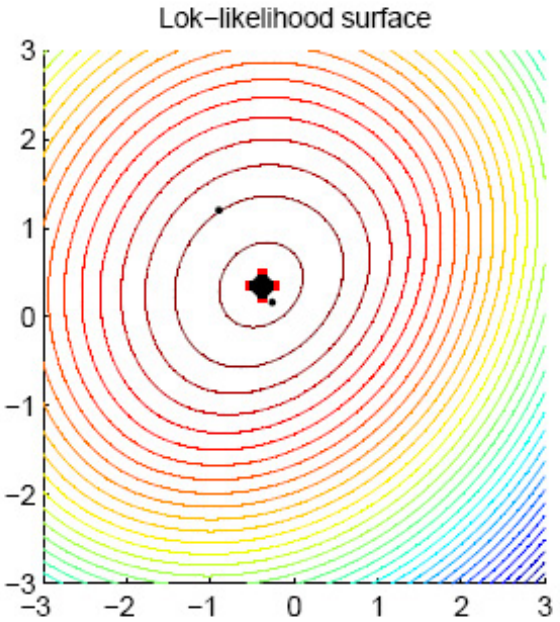
- This is our objective function, and we can find its peak by gradient descent as before.
  - Need to modify the calculation of gradient and Hessian.

Regularized logistic regression – demo

Regularized logistic regression – demo

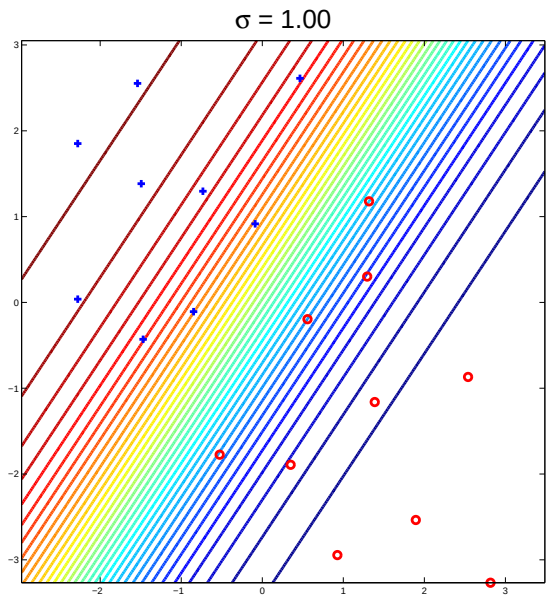




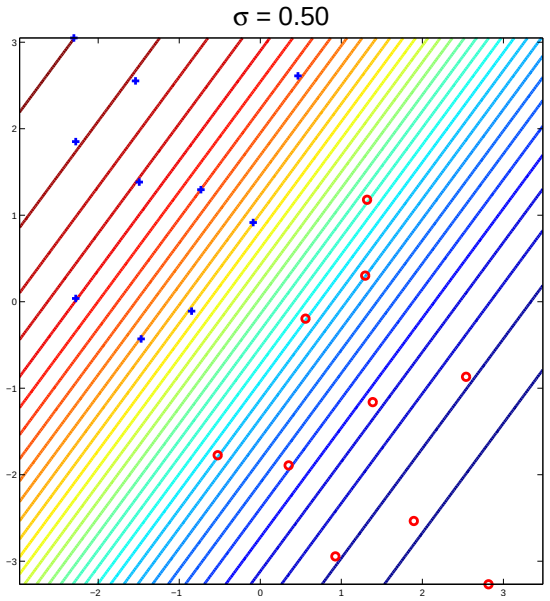


The effect of regularization: separable data

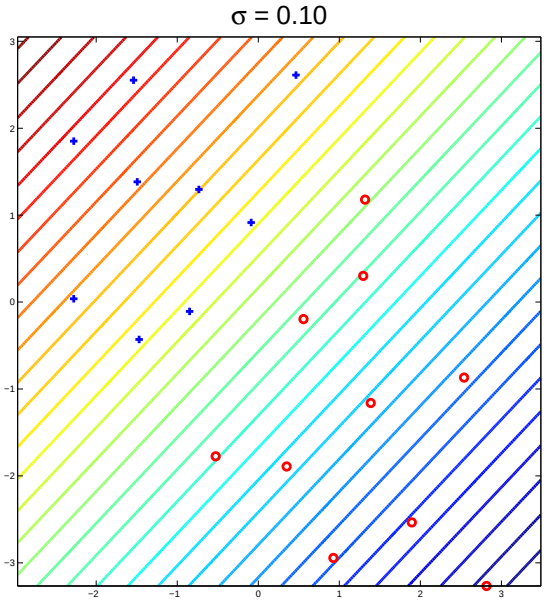
$$\tilde{\ell}(X_N; \mathbf{w}, \sigma) \;=\; \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}\right) \;-\; \frac{1}{2\sigma^2}\|\mathbf{w}\| \;+ \text{const}(\mathbf{w}).$$



$\sigma^2 = 1$



$\sigma^2 = 0.5$

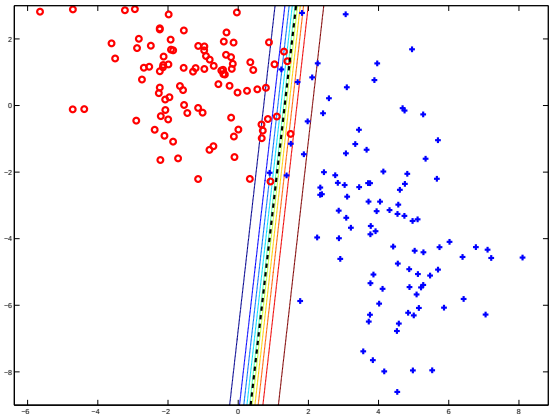


$\sigma^2 = 0.1$

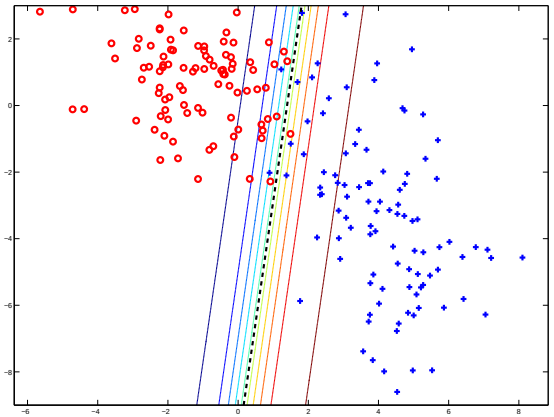
The effect of regularization

$$\tilde{\ell}(X_N; \mathbf{w}, \sigma) \;=\; \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}\right) \;-\; \frac{1}{2\sigma^2}\|\mathbf{w}\| \;+\; \text{const}(\mathbf{w}).$$

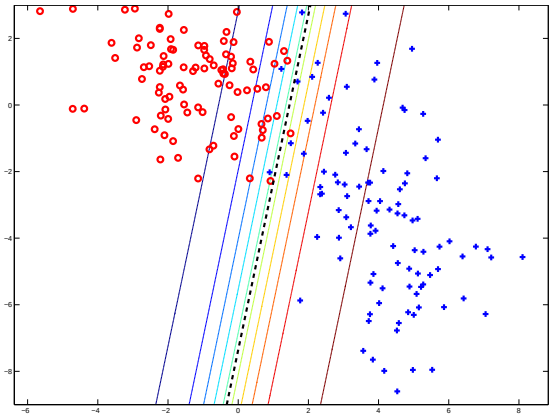
ML



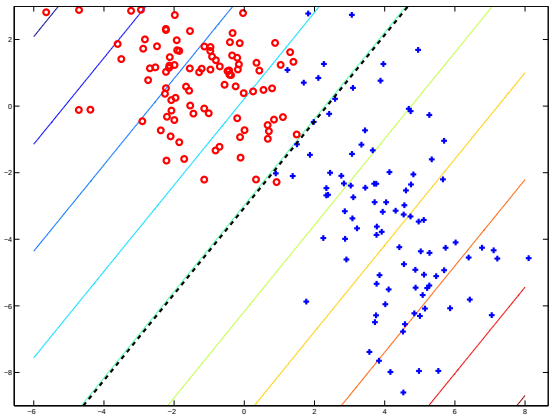
$\sigma^2 = 1$



$\sigma^2 = 0.1$



$\sigma^2 = 0.01$



Scaled objective

- When N is large, it may be more convenient to work with

$$\tilde{\ell}(X_N; \mathbf{w}, \sigma) \;=\; \frac{1}{N} \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}\right) \;-\; \frac{1}{2N\sigma^2} \|\mathbf{w}\|^2.$$

- The parameter $\lambda = \frac{1}{2\sigma^2}$ determines the *strength* of regularization.
  - Higher $\lambda \Rightarrow$ more weight on the prior.

L_2 versus L_1 regularization

- We can have a different penalty term. For instance,

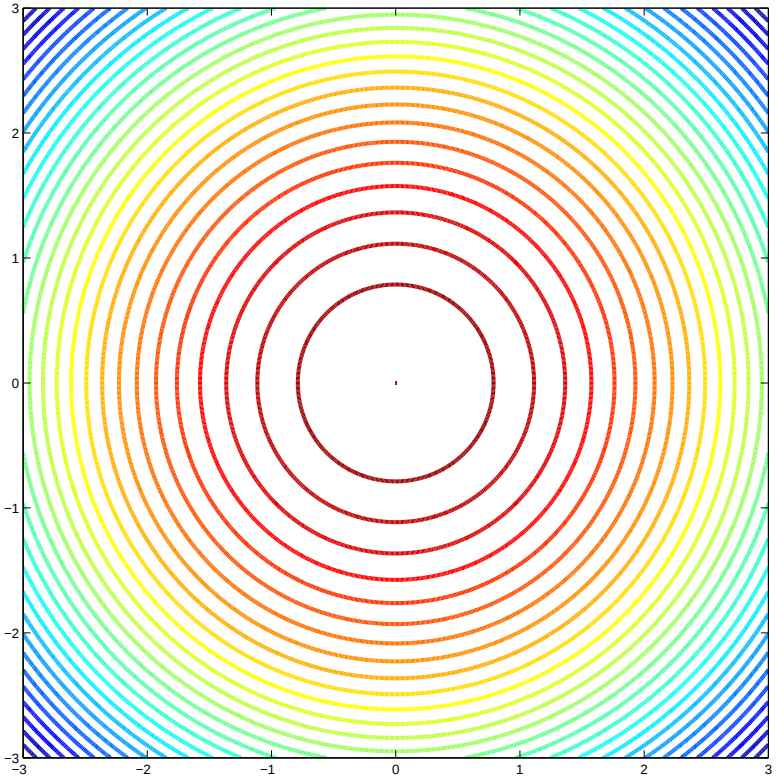
$$\tilde{\ell}(X_N; \mathbf{w}, \sigma) = \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}\right) - \lambda \sum_{j=1}^{d} |w_j| + \text{const}(\mathbf{w}),$$

  the L_1-norm penalty.

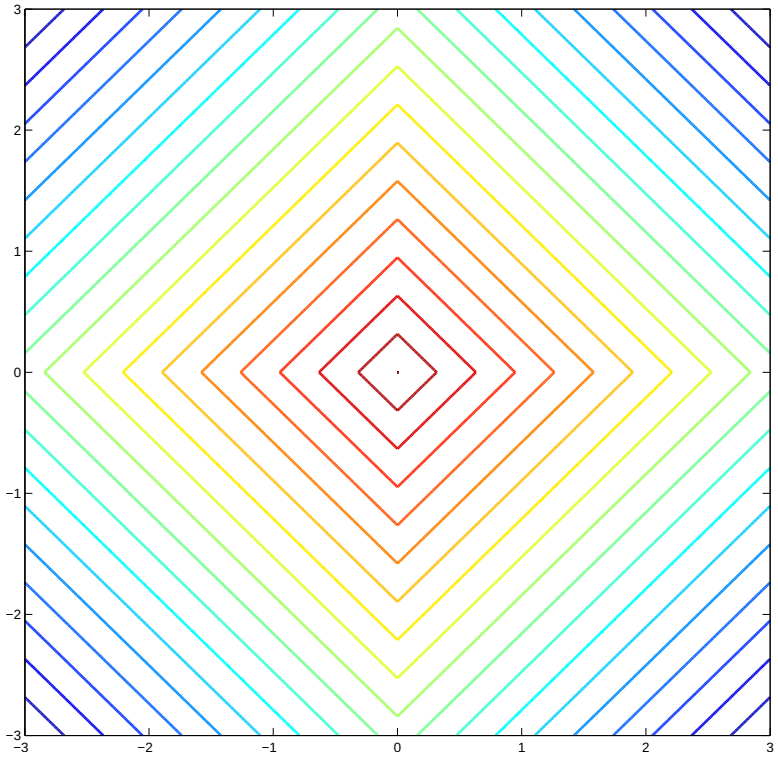
  - the L_p norm of $\mathbf{w}$ is $\left(\sum_{j=1}^{d} w_j^p\right)^{1/p}$.
  - L_2: Euclidean distance.
  - L_1: Manhattan distance.

L_2 versus L_1 regularization

$-\lambda \sum_{j=1}^{d} w_j^2$



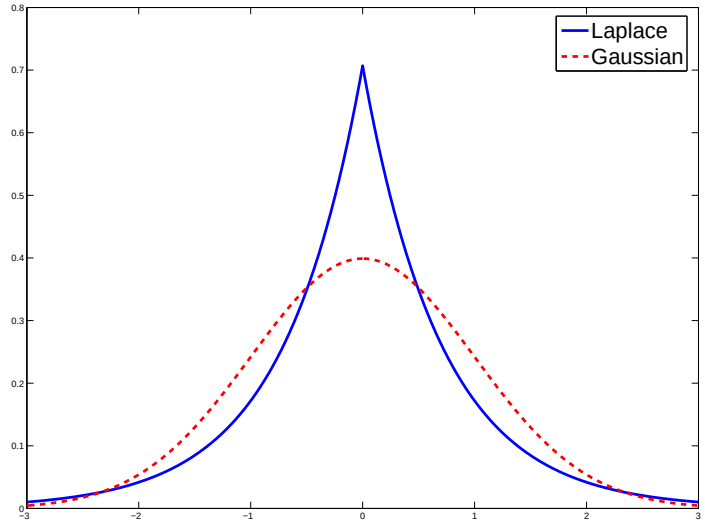
$-\lambda \sum_{j=1}^{d} |w_j|$



Probabilistic interpretation of L_1 penalty

- Laplacian pdf:

$$p(x; a) = \frac{a}{2} e^{-a|x|}$$



- L_1 drives $\mathbf{w}$ to zero more efficiently.

Next time

Regularization for regression.
Large-margin classification.