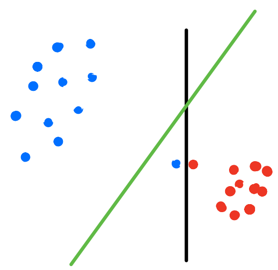
Recap Separating Hyperplanes
Hard-Margin and Soft-Margin SVM

Today Least Squares Regression



$$\min_{w,b} \frac{1}{2}\|w\|_2^2$$
$$y_i(w^T x_i + b) \geq 1 \quad \forall i$$

Hard-Margin SVM

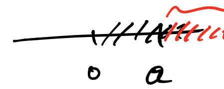
$$\min \frac{1}{2}\|w\|_2^2 + C \cdot \frac{1}{n}\sum_{i=1}^{n} s_i$$
$$y_i(w^T x_i + b) \geq 1 - s_i \quad \forall i$$
$$s_i \geq 0 \quad \forall i$$

Soft-Margin SVM

$$\Downarrow$$
$$s_i \geq \max(0, 1 - y_i(w^T x_i + b))$$
$$= (1 - y_i(w^T x_i + b))_+$$

$s_i \geq a$

$s_i \geq 0$



$s \geq \max(0, a)$

$\max(0, a) = (a)_+$

$$\min_{\underline{w}, b} \frac{1}{2}\|w\|_2^2 + C\frac{1}{n}\sum_{i=1}^{n}(1 - y_i(w^T x_i + b))_+$$

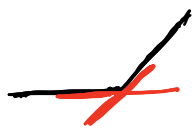
Soft-Margin SVM.

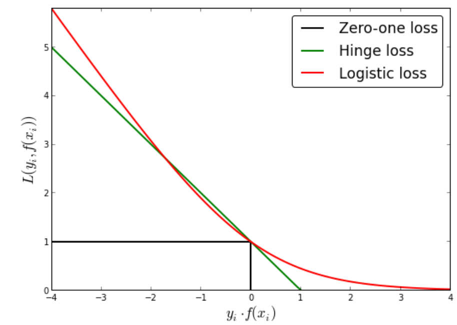
$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))$$

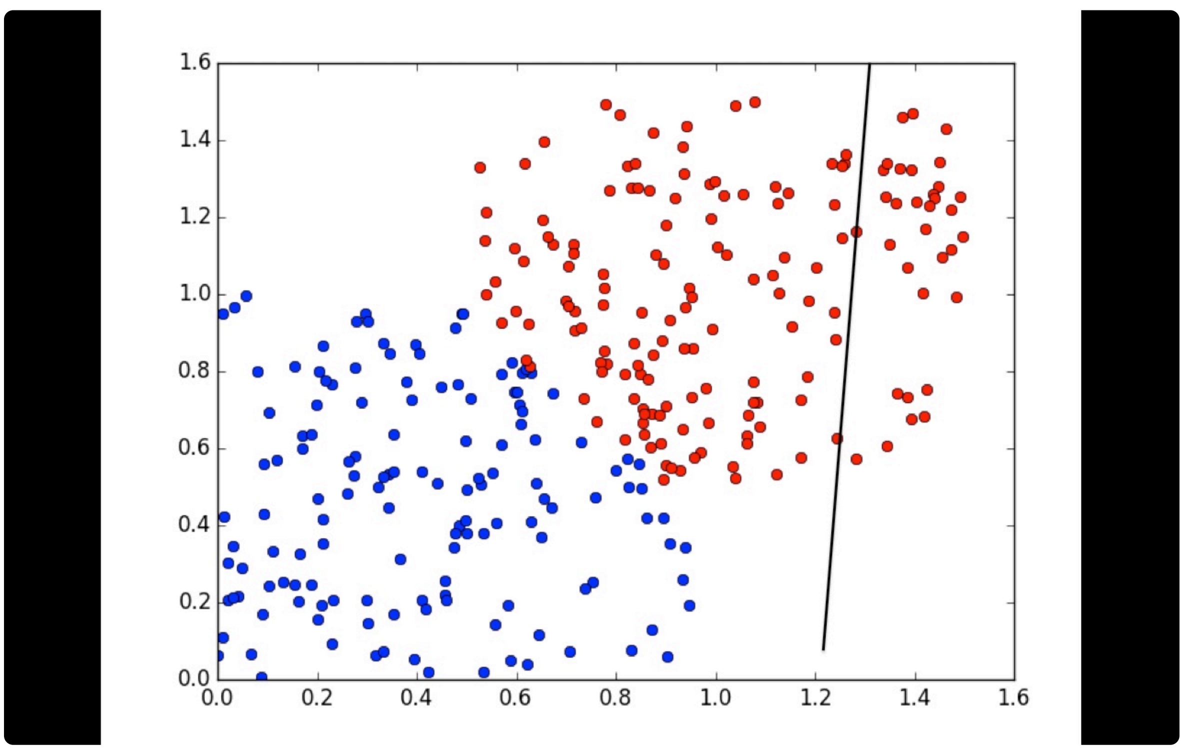
- Hinge Loss: $L(y, f(x)) = (1 - y f(x))_+$ where $f(x) = wx + b$
  (Soft-Margin SVM)

- Logistic Loss $L(y, f(x)) = \log(1 + e^{-yf(x)})$
  (Logistic Regression)

We can use Gradient Descent (sub-gradient Descent)







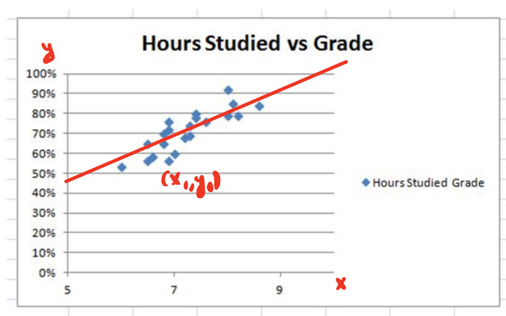
Least Squares Regression

Predict the value of a continuous target variable y

$(x_1, y_1), \ldots, (x_n, y_n)$ training data

$X \in \mathbb{R}^d$, $Y \in \mathbb{R}$ random.

Linear Regression $f(x) = w^T x + b$



Performance measure mean squared error (MSE)

$$R(f) = \mathbb{E}_{X,Y}\left[(Y - f(X))^2\right]$$

$f(x)$ linear $f(x) = w^T x + b$

$$R(w, b) = \mathbb{E}_{X,Y}\left[(Y - w^T X - b)^2\right]$$

$P_{X,Y}$ unknown. We can estimate R directly

$$\hat{R}(w, b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i - b)^2$$

Add a regularization term $\|w\|_2^2$

$$\min_{w,b} \overbrace{\frac{1}{n}\sum (y_i - w^T x_i - b)^2 + \lambda \|w\|_2^2}^{L(w,b)}$$

$\lambda = 0 \Rightarrow$ LS Regression

$\lambda > 0 \Rightarrow$ Ridge Regression

$$\frac{\partial}{\partial b} L(w,b) = 2\frac{1}{n}\sum (y_i - w^T x_i - b)\cdot(-1) = 0$$

$$b^* = \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i) = \bar{y} - w^T\bar{x}$$

where $\bar{y} = \frac{1}{n}\sum y_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Plugging b^* in L

$$\min_{w} \frac{1}{n}\sum (y_i - \bar{y} - w^T(x_i - \bar{x}))^2 + \lambda\|w\|_2^2$$

Define $\tilde{y}_i = y_i - \bar{y} \qquad \tilde{x}_i = x_i - \bar{x}$

$$\min \frac{1}{n}\|\tilde{y} - \tilde{X}w\|_2^2 + \lambda\|w\|_2^2 \qquad \tilde{y} = \begin{bmatrix}\tilde{y}_1\\ \vdots \\ \tilde{y}_n\end{bmatrix} \quad \tilde{X} = \begin{bmatrix}\tilde{x}_1^T\\ \vdots \\ \tilde{x}_n^T\end{bmatrix}$$

$$w^* = (\tilde{X}^T\tilde{X} + n\lambda I)^{-1}\tilde{X}^T\tilde{y} \qquad b^* = \bar{y} - (w^*)^T\bar{x}$$

Directly: $O(nd^2)$

→ Gradient Descent $\quad w_{t+1} = w_t - \alpha_t \cdot 2((\tilde{X}^T\tilde{X} + n\lambda I)w_t - \tilde{X}^T\tilde{y})$

One iteration: $O(nd)$

$\tilde{X}^T(\tilde{X} w_t)$, $\tilde{X}^T\tilde{y}$

$$= w_t - \alpha_t \sum_{i=1}^{n} 2\tilde{x}_i(w^T\tilde{x}_i - \tilde{y}_i) + 2\lambda w_t$$

- n is large
- dataset does not fit into memory.

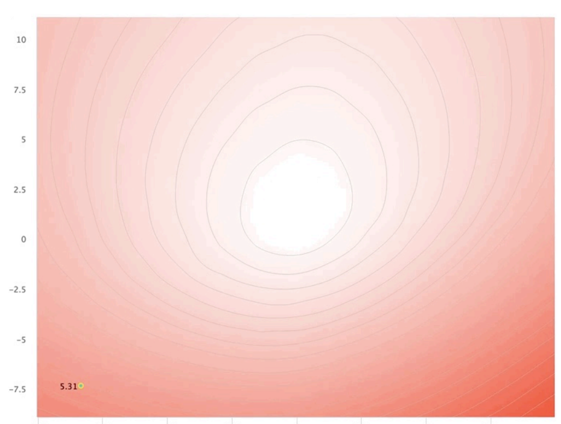
→ Stochastic Gradient Descent

For $i=1,\ldots,n$ in random order

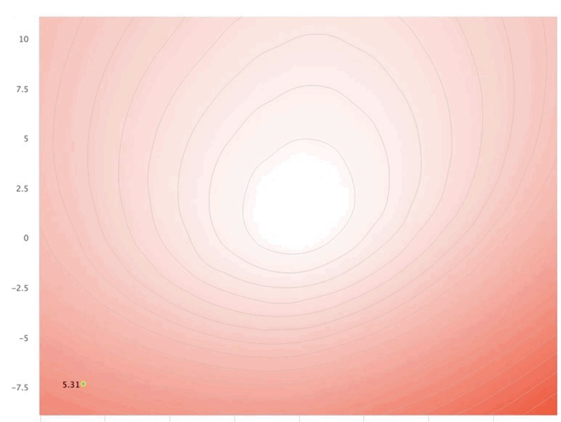
$$w_{t+1} = w_t - 2\alpha_t\left(2\tilde{x}_i(w^T\tilde{x}_i - \tilde{y}_i) + 2\lambda w_t\right)$$

One iteration: $O(d)$

- n is extremely large
- online data



GD



SGD

Probability Model: Suppose

$$p(y \mid x,\theta) = N(w^Tx + b, \sigma^2) \Rightarrow E[y_i \mid x_i] = w^Tx_i + b$$

$$\theta = \begin{bmatrix} w \\ b \end{bmatrix}$$

$$= \frac{1}{\quad} \cdot e^{-\frac{(y - w^Tx - b)^2}{2\sigma^2}}$$

$\sqrt{2\pi\sigma^2}$

$$\arg \max_{\theta} \log \prod_{i=1}^{n} P(y_i | x_i, \theta)$$

$$\theta_{ML} = \arg \max_{\theta} -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i - b)^2 - \frac{n}{2} \log 2\pi\sigma^2$$

Maximum Likelihood Estimate = Linear Least Squares

MAP estimation: $p(y | x, \theta) = N(w^T x + b, \sigma^2)$

Suppose we have a prior dist on w

$$p(w) = N(0, t^2 I) \propto e^{\frac{-\|w\|_2^2}{2t^2}}$$

$$\hat{w}_{MAP} = \arg \max_{w} \log P(y | x \theta) \cdot P(w)$$

$$= \arg \max_{w} -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i - b)^2 - \frac{1}{2t^2} \|w\|_2^2$$

$$= \arg \min \sum_{i=1}^{n} (y_i - w^T x_i - b)^2 + \underbrace{\frac{\sigma^2}{t^2}}_{= \lambda} \cdot \|w\|_2^2$$

$$\arg \min \sum_{i=1}^{n} (y_i - w^T x_i - b)^2 + \lambda \cdot \|w\|_p$$