Gradient Descent

Natural Language Processing
Emory University
Jinho D. Choi

Supervised Learning

input

output $y = \pm 1$ binomial distribution

$$(X, Y) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

prediction $\hat{y} = f(x)$ predicts the output of x

Expected risk $E(f) = \int \ell(\hat{y}; y) \cdot P(x, y)$

loss function

joint distribution unknown!

Empirical risk $\hat{E}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{y}_i; y_i)$ minimize!

Linear Prediction

$$\hat{E}(f) = \frac{1}{n} \sum_{i=1}^{n} \boxed{\ell(\hat{y}_i; y_i)}$$

$$\ell(\hat{y}; y) = \boxed{\frac{1}{2}(\hat{y} - y)^2}$$ least squares

$$\hat{y} = f(x) = \mathrm{w}^T \boxed{\Phi(x)} = \boxed{\mathrm{w}^T \mathrm{x}}$$ linear function

feature vector

$$\boxed{\ell(\mathrm{w}, x; y)} = \frac{1}{2}(\boxed{\mathrm{w}^T}\mathrm{x} - y)^2$$

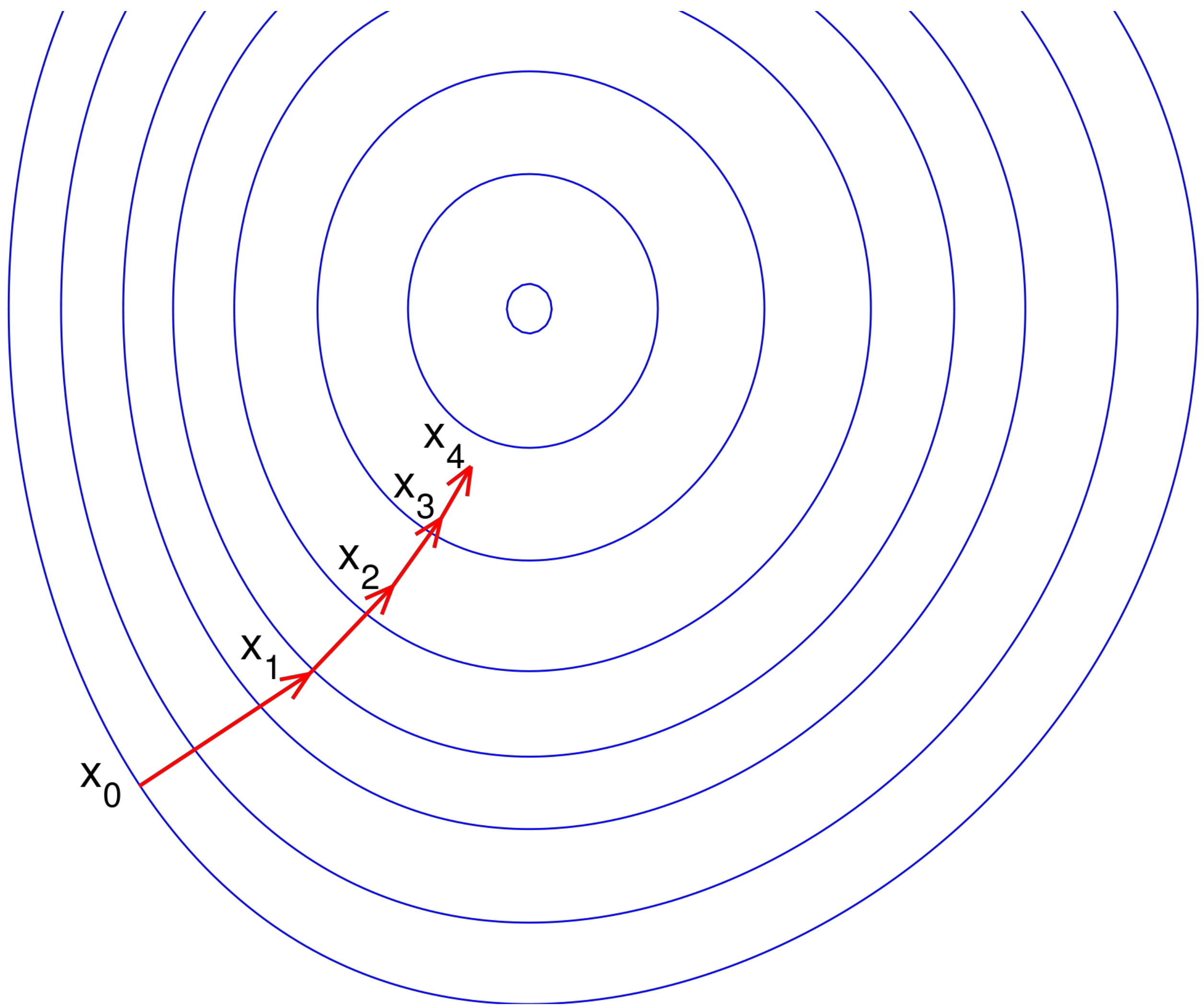
Find a weight vector that minimizes the loss.

Gradient Descent

learning rate

derivative of the loss

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t - \boxed{\eta_t} \frac{1}{n} \sum_{i=1}^{n} \boxed{\frac{\partial}{\partial \mathrm{w}} \ell(\mathrm{w}_t, x_i; y_i)}$$



Minimize loss

Derivative → 0

Convex optimization

Global optimum?

Gradient Descent

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t - \eta_t \frac{1}{n} \sum_{i=1}^{n} \boxed{\frac{\partial}{\partial \mathrm{w}} \boxed{\ell(\mathrm{w}_t, x_i; y_i)}}$$

$$\ell(\mathrm{w}, x; y) = \frac{1}{2}(\mathrm{w}^T \mathrm{x} - y)^2$$

$$\frac{\partial}{\partial w} \ell(\mathrm{w}, x; y) = \frac{\partial}{\partial \mathrm{w}} \frac{1}{2}(\mathrm{w}^T \mathrm{x} - y)^2 = \boxed{(\mathrm{w}^T \mathrm{x} - y)\mathrm{x}}$$

$$\boxed{\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t} - \eta_t \frac{1}{n} \sum_{i=1}^{n} (\mathrm{w}^T \mathrm{x}_i - y_i)\mathrm{x}_i$$

How often is the weight vector updated?

Stochastic Gradient Descent

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t - \eta_t \boxed{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{w}^T\mathrm{x}_i - y_i)\mathrm{x}_i}$$

$$\boxed{\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t} - \eta_t \boxed{(\mathrm{w}_t^T\mathrm{x}_i - y_i)\mathrm{x}_i}$$

updated for every instance

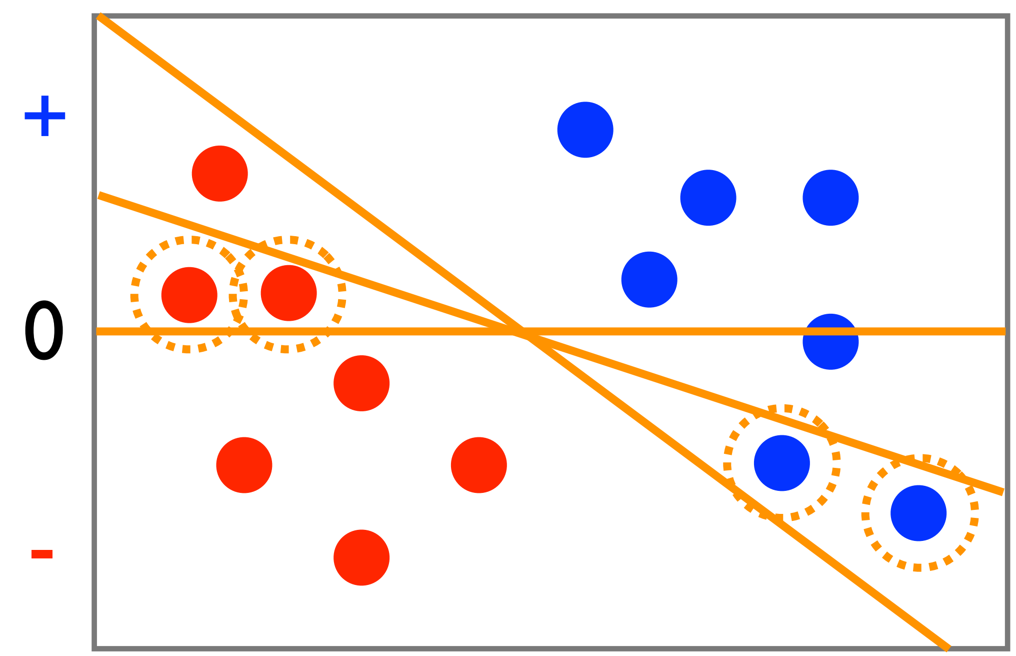
$$\mathrm{w}_0 \leftarrow 0$$

$$\mathrm{w}_0^T\mathrm{x}_1 > 0 \quad \mathrm{w}_1 \leftarrow \mathrm{w}_0 - \eta(\oplus + 1)\mathrm{x}_1$$

$$\mathrm{w}_1^T\mathrm{x}_2 < 0 \quad \mathrm{w}_2 \leftarrow \mathrm{w}_1 - \eta(\ominus + 1)\mathrm{x}_2$$

$$\mathrm{w}_2^T\mathrm{x}_3 < 0 \quad \mathrm{w}_3 \leftarrow \mathrm{w}_2 - \eta(\ominus - 1)\mathrm{x}_3$$

$$\mathrm{w}_3^T\mathrm{x}_4 > 0 \quad \mathrm{w}_4 \leftarrow \mathrm{w}_3 - \eta(\oplus - 1)\mathrm{x}_4$$



Perceptron

Stochastic gradient descent

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t - \eta_t \Delta\ell$$

Least squares

$$\ell(\mathrm{w}, x; y) = \frac{1}{2}(\mathrm{w}^T\mathrm{x} - y)^2$$

$$\Delta\ell = (\mathrm{w}^T\mathrm{x} - y)\mathrm{x}$$

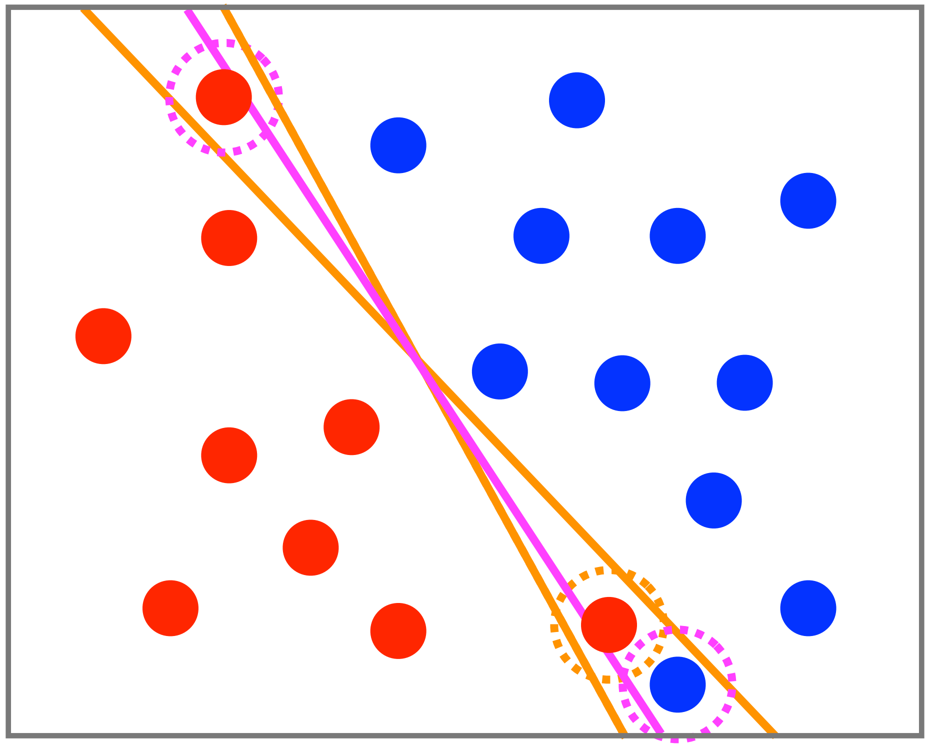
Perceptron

$$\ell(\mathrm{w}, \mathrm{x}; y) = \boxed{\max\{0, -\mathrm{w}^T\mathrm{x} \cdot y\}}$$

$$\Delta\ell = \begin{cases} -\mathrm{x} \cdot y & \mathrm{w}^T\mathrm{x} \cdot y < 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + \eta_t \begin{cases} \mathrm{x} \cdot y & \mathrm{w}_t^T\mathrm{x} \cdot y < 0 \\ 0 & \text{otherwise} \end{cases}$$

Averaged Perceptron



The final hyperplane may be overfitted to later instances.

Take the average of all hyperplanes including ones that are not updated.

Averaged Perceptron

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + \eta_t(\mathrm{x} \cdot y) \quad \text{if } \mathrm{w}_t^T \boxed{\mathrm{x}} \cdot y < 0$$

sparse vector?

$$\overline{\mathrm{w}} \leftarrow \frac{1}{c} \sum_{t=0}^{c-1} \mathrm{w}_t$$

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + \eta_t(\mathrm{x} \cdot y)$$

$$\mathrm{v}_{t+1} \leftarrow \mathrm{v}_t + \eta_t \cdot \boxed{c}(\mathrm{x} \cdot y)$$

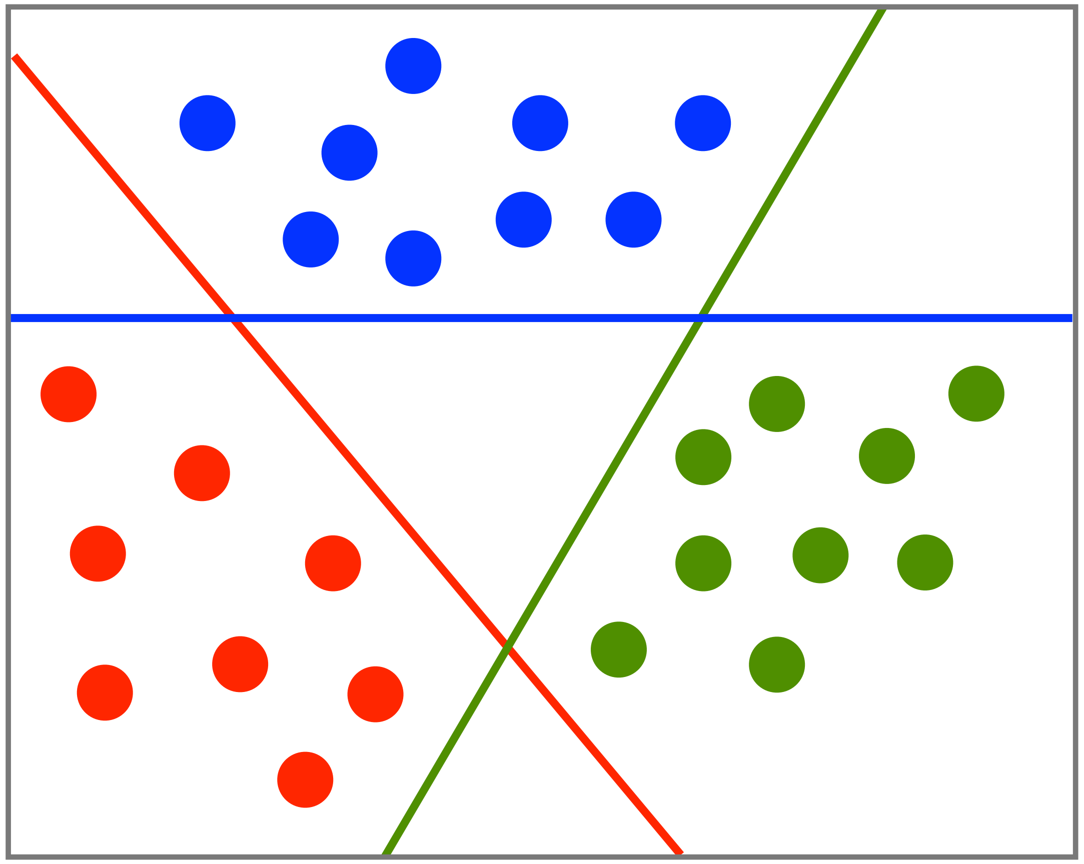
Initialization: $c \leftarrow 1$

Update rule: $c \leftarrow c + 1$ for every instance

$$\overline{\mathrm{w}} \leftarrow \mathrm{w} - \boxed{\frac{1}{c} \cdot \mathrm{v}}$$

Multinomial Perceptron

Binomial distribution requires
1 hyperplane to separate 2 classes.



How many for
m classes?

Multinomial distribution requires
m hyperplanes to separate *m* classes.

Multinomial Perceptron

$$\mathrm{x} = \begin{array}{|c|c|c|c|c|}\hline \boxed{1} & 0 & 0 & \boxed{1} & 0 \\ \hline\end{array}$$

5 features (including bias)

Binomial $y = \{-1, 1\}$

$$\mathrm{w} = \begin{array}{|c|c|c|c|c|}\hline \boxed{a} & b & c & \boxed{d} & e \\ \hline\end{array}$$

$$\mathrm{w}^T\mathrm{x} = a + d \qquad \hat{y} = \begin{cases} 1 & \mathrm{w}^T\mathrm{x} \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

Multinomial $y = \{0, 1, 2, 3\}$

$$\mathrm{w} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}\hline a_0 & a_1 & a_2 & a_3 & b_0 & b_1 & b_2 & b_3 & c_0 & c_1 & c_2 & c_3 & d_0 & d_1 & d_2 & d_3 & e_0 & e_1 & e_2 & e_3 \\ \hline\end{array}$$

$$\mathrm{w}_y^T\mathrm{x} = a_y + d_y \qquad \hat{y} = \arg\max_y \mathrm{w}_y^T\mathrm{x}$$

Binomial vs. Multinomial Perceptron

$$\text{if } \mathrm{w}_t^T \mathrm{x} \cdot y < 0 \Leftrightarrow y \neq \hat{y}$$

Binomial

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + \eta_t(\mathrm{x} \cdot y)$$

Multinomial

$$\mathrm{w}_{y,t+1} \leftarrow \mathrm{w}_{y,t} + \eta_t \cdot \mathrm{x}$$
$$\mathrm{w}_{\hat{y},t+1} \leftarrow \mathrm{w}_{\hat{y},t} - \eta_t \cdot \mathrm{x}$$

Hinge Loss

Perceptron

$$\ell(\mathrm{w}, \mathrm{x}; y) = \boxed{\max\{0, -\mathrm{w}^T \mathrm{x} \cdot y\}}$$

$$\Delta\ell = \begin{cases} -\mathrm{x} \cdot y & \boxed{\mathrm{w}^T \mathrm{x} \cdot y < 0} \\ 0 & \text{otherwise} \end{cases}$$

Hinge loss

$$\ell(\mathrm{w}, \mathrm{x}; y) = \boxed{\max\{0, 1 - \mathrm{w}^T \mathrm{x} \cdot y\}}$$

$$\Delta\ell = \begin{cases} -\mathrm{x} \cdot y & \boxed{\mathrm{w}^T \mathrm{x} \cdot y < 1} \\ 0 & \text{otherwise} \end{cases}$$

Adaptive Gradient Descent

Perceptron

$$\text{if } \mathrm{w}_t^T \mathrm{x} \cdot y < 0$$

Hinge loss

$$\text{if } \mathrm{w}_t^T \cdot y < 1$$

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + \boxed{\eta_t}(\mathrm{x} \cdot y)$$

$$\mathrm{g}_{t+1} \leftarrow \mathrm{g}_t + \mathrm{x} \circ \mathrm{x}$$

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + \boxed{\frac{\eta}{\rho + \sqrt{g_{t+1}}}} \cdot (\mathrm{x} \cdot y)$$