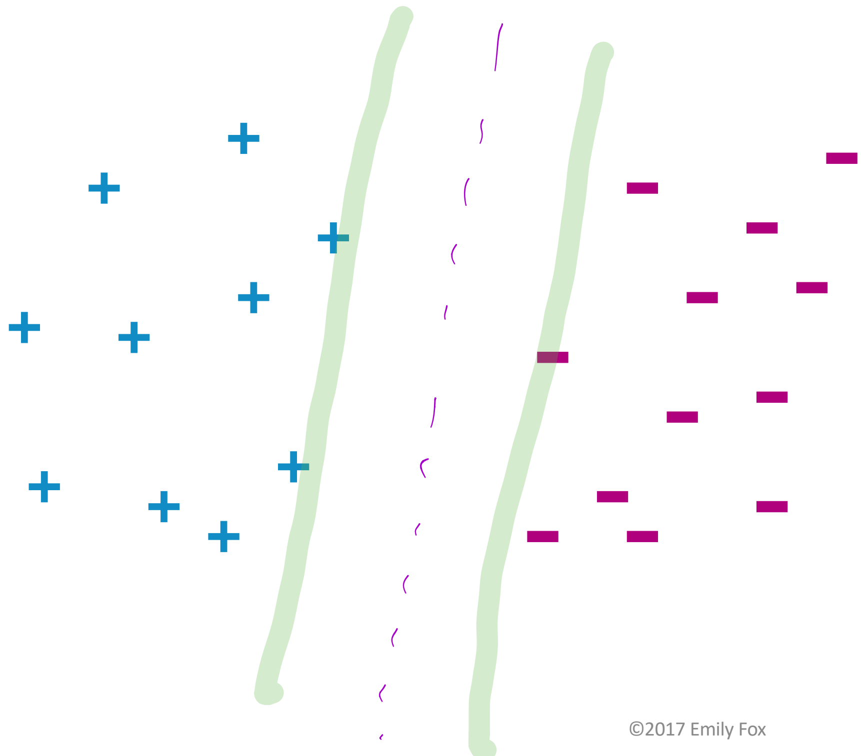
Support Vector Machines

CSE 446: Machine Learning
Slides by Emily Fox + Kevin Jamieson + others

Presented by Anna Karlin

May 1, 2019

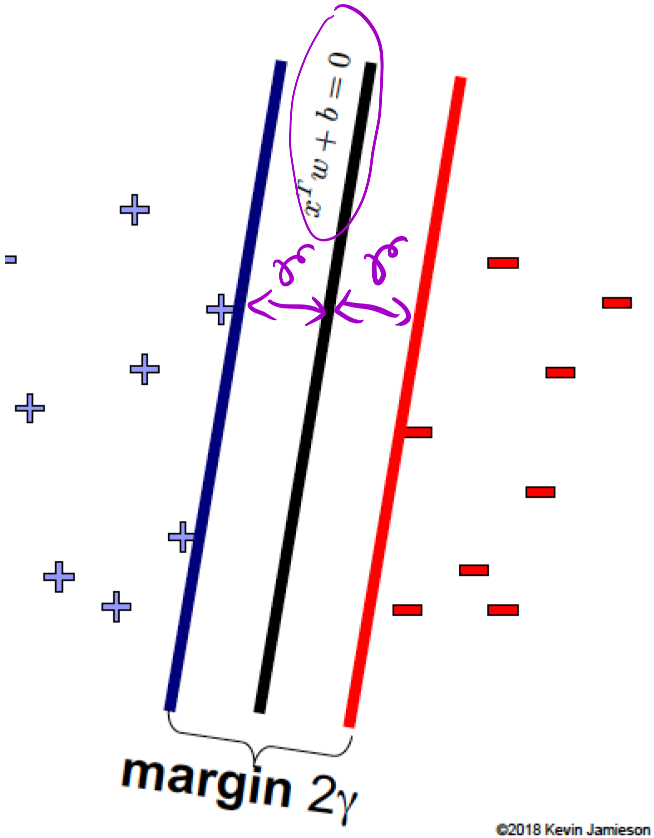
Linear classifiers—Which line is better?



$\{(x_i, y_i)\}$ $x_i \in \mathbb{R}^d$

$y_i \in \{-1, +1\}$

find linear decision boundary with the largest margin possible



find w & b
that gives largest γ

Maximizing the margin for linearly separable data

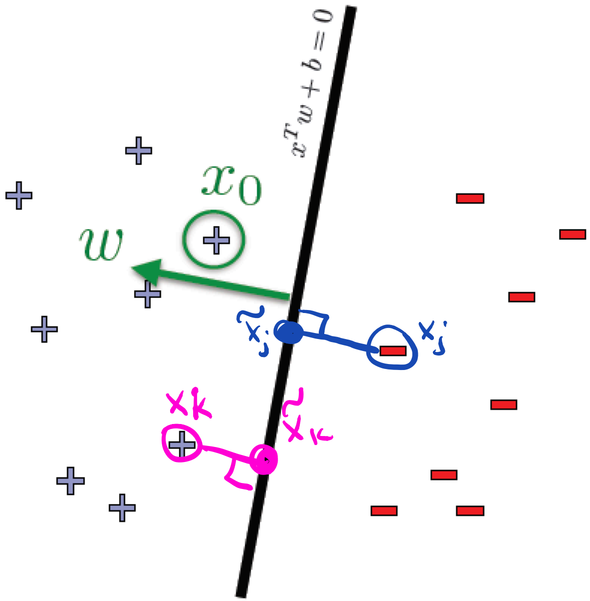
$\frac{w}{\|w\|}$ unit vector

$u \cdot v = \|u\| \|v\| \cos\theta$

$\tilde{x}_k \cdot w + b = 0$

$\tilde{x}_j \cdot w + b = 0$

Given hyperplane, what is margin?



Distance from x_0 to hyperplane defined by $x^T w + b = 0$?

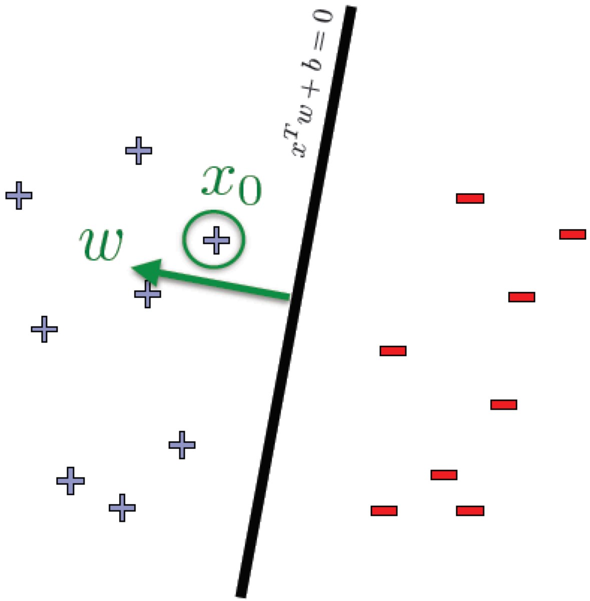
$y_k = +1 \qquad \|x_k - \tilde{x}_k\| = (x_k - \tilde{x}_k) \cdot \frac{w}{\|w\|} = \frac{x_k \cdot w + b}{\|w\|} \geq \gamma$

$y_j = -1 \qquad (x_j - \tilde{x}_j) \cdot \frac{w}{\|w\|} = -\|x_j - \tilde{x}_j\|$

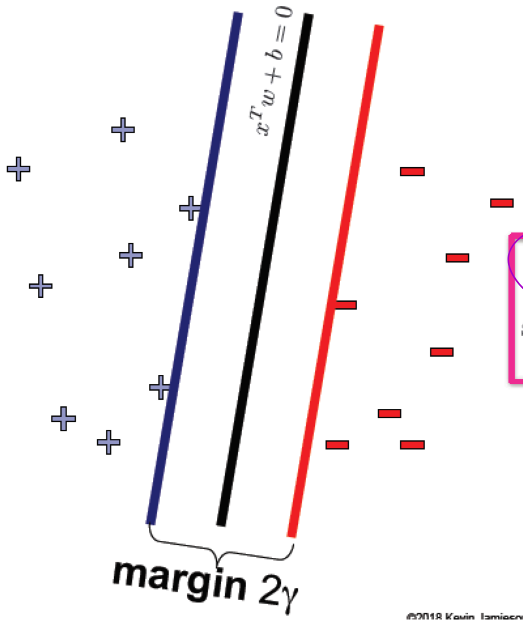
want $\quad -\frac{(x_j \cdot w + b)}{\|w\|} \geq \gamma$

$\forall\, 1 \leq i \leq n \qquad y_i \left(\frac{x_i \cdot w + b}{\|w\|}\right) \geq \gamma$

Given hyperplane, what is margin?



Our optimization problem



Optimal Hyperplane

$$\max_{w,b} \gamma$$

$$\underbrace{\text{subject to } \frac{1}{||w||_2} y_i(x_i^T w + b) \geq \gamma}_{100w \quad 100b} \quad \forall i$$

$y_i(x_i^T w + b) \geq \gamma ||w||_2$

$100w \quad 100b \quad 100w$

Choose it so that $\boxed{\gamma ||w||_2 = 1}$

$\max \gamma$

s.t. $y_i(x_i^T w + b) \geq 1$

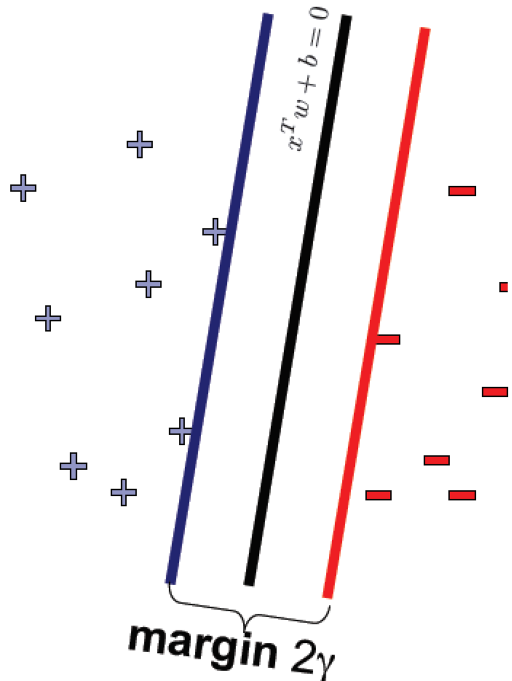
where $\gamma = \frac{1}{\|w\|_2}$

$\equiv \max \frac{1}{\|w\|_2} \equiv \min \|w\|_2$

$\equiv \min \|w\|_2^2$

Final version

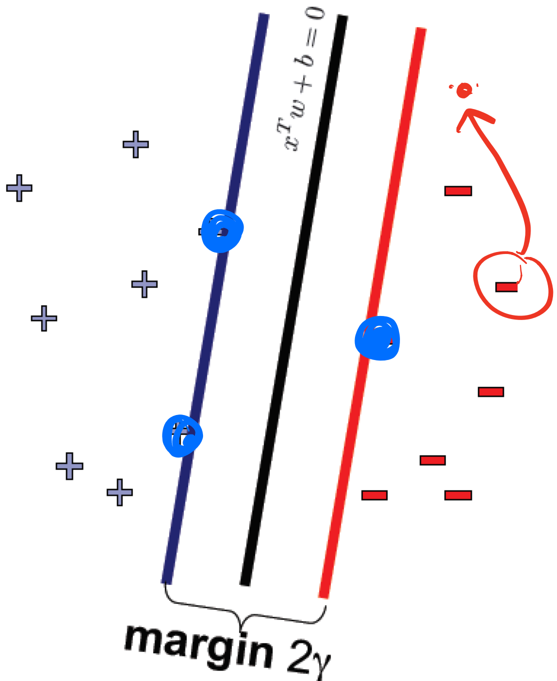
Solvable efficiently – quadratic programming problem



Optimal Hyperplane (reparameterized)

$$\boxed{\begin{array}{l} \min\limits_{w,b} ||w||_2^2 \\ \text{subject to } y_i(x_i^T w + b) \geq 1 \quad \forall i \end{array}}$$

What are support vectors?



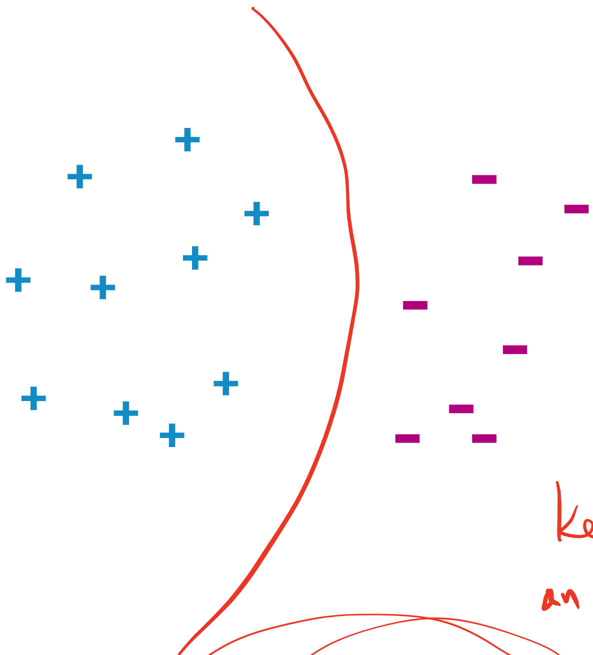
Optimal Hyperplane (reparameterized)

$$\min_{w,b} ||w||_2^2$$

$$\text{subject to } y_i(x_i^T w + b) \geq 1 \quad \forall i$$

What if the data are not linearly separable?

What if data are not linearly separable?



Use feature maps…

$\phi : \mathbb{R}^d \to F \quad (\mathbb{R}^p)$

solve problem on $\{(\phi(x_i), y_i)\}_{i=1}^n$

SVMs can be kernelized!!!!

write alg in terms of inner products

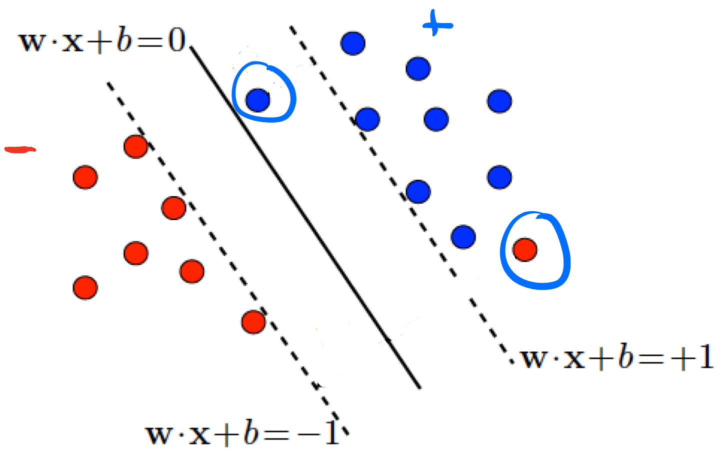
$x_i \cdot x_j \Longrightarrow \phi(x_i) \cdot \phi(x_j)$

Kernel fns $K(x, x')$ that implements an inner product $= \phi(x) \cdot \phi(x')$

$w = \sum_{i=1}^{n} \alpha_i y_i x_i$

$\alpha_i \geq 0$

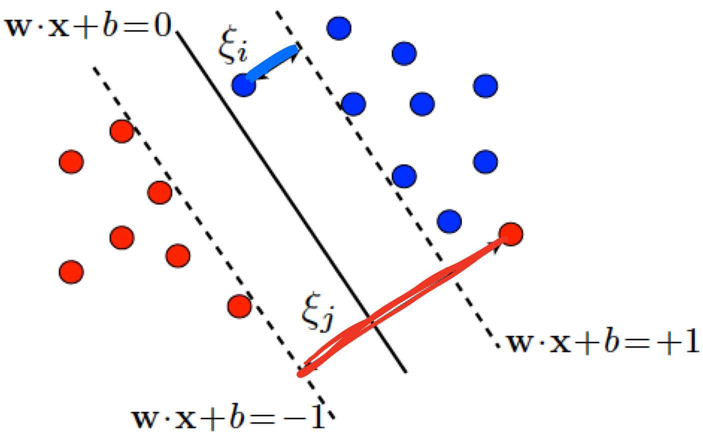
What if data are still not linearly separable?



Courtesy Mehryar Mohri

What if data are still not linearly separable?

allow for slack



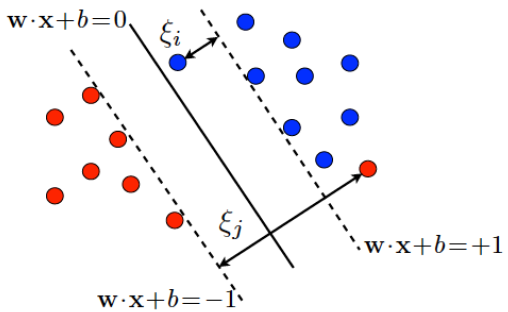
Courtesy Mehryar Mohri

$\min \|w\|_2^2$

s.t. $y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1-\xi_i \quad \forall i$

$\xi_i \geq 0$

What if data are still not linearly separable?



Courtesy Mehryar Mohri

total slack

$$\min_{\mathbf{w},b,\xi} \; \boxed{||\mathbf{w}||^2 + C\sum_i \xi_i}$$

$$\boxed{y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i} \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i.$$

Fix $\vec{w}, b$

$y_j(w \cdot x_j + b) \geq 1 \implies \xi_j = 0$

$y_k(w \cdot x_k + b) < 1 \implies \xi_k = 1 - y_k(w \cdot x_k + b)$

Replace obj fn with

$$\min_{w,b} \; ||w||^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(w \cdot x_i + b))$$

$$z_+ = \max(z, 0)$$

Final objective

$\lambda = \frac{1}{nC}$

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{\boxed{(1 - y_i((\mathbf{w}^T\mathbf{x}_i + b))_+}}_{\text{loss on } i^{th} \text{ data pt}} + \lambda ||\mathbf{w}||_2^2$$

hinge loss

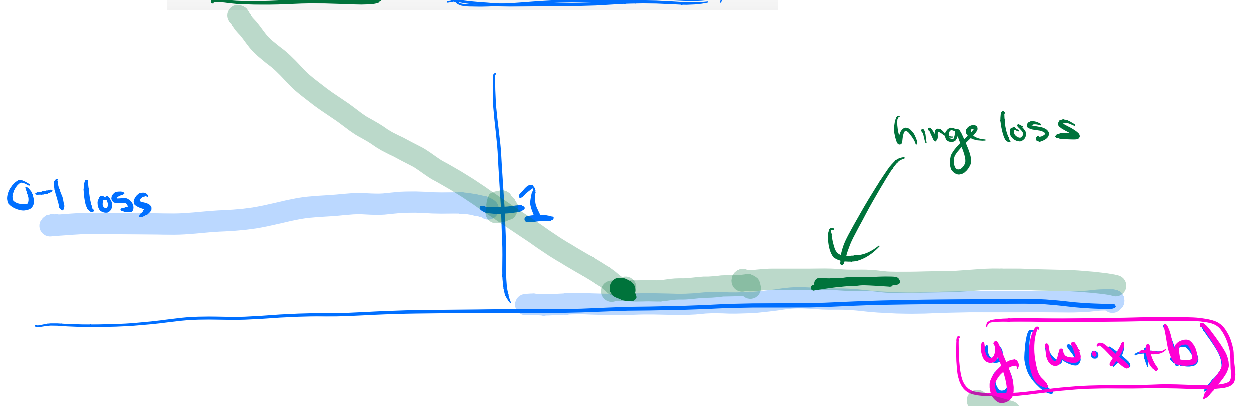
Gradient descent for SVMs

0-1 loss on (x,y) $\boxed{\mathbb{1}\{\text{sign}(w^Tx+b) \neq y\}}$ mistake on this pt

Hinge loss

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i((\mathbf{w}^T\mathbf{x}_i + b))_+ + \lambda||\mathbf{w}||_2^2$$

- Hinge loss: $\ell((\mathbf{x}, y), \mathbf{w}) = (1 - y(\mathbf{w}^T\mathbf{x} + b))_+$



measures correctness of our prediction

- Subgradient of hinge loss:

d=1

$$\partial \ell_w = \begin{cases} 0 & y(wx+b) > 1 \\ [-yx, 0] & y(w\cdot x+b) = 1 \\ -yx & 1 - y(wx+b) > 0 \end{cases}$$

Minimizing regularized hinge loss (aka SVMs)

- Given a dataset: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$
- Minimize regularized hinge loss: $\frac{1}{n}\sum_i (1 - y_i((\mathbf{w}^T\mathbf{x}_i + b))_+ + \lambda||\mathbf{w}||_2^2$

Subgradient descent for hinge minimization

- Given data: $$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$$
- Want to minimize:

$$\frac{1}{n}\sum_{i=1}^{n} \ell((\mathbf{x}_i, y_i), \mathbf{w}) + \lambda||\mathbf{w}||_2^2 = \frac{1}{n}\sum_{i}(1 - y_i((\mathbf{w}^T\mathbf{x}_i + b))_+ + \lambda||\mathbf{w}||_2^2$$

- As we've discussed, subgradient descent works like gradient descent:
  - But if there are multiple subgradients at a point, just pick (any) one:

$$\partial_{\mathbf{w}}\ell((\mathbf{x}, y), \mathbf{w}) = \mathbb{I}\{y(\mathbf{w}^T\mathbf{x} + b) \leq 1\}(-y\mathbf{x})$$

subgradient. $\vec{g}$ is a subgradient at w for $f(\ )$

$f(w') \geq f(w) + \vec{g}\cdot(w' - w)$

Subgradient descent for hinge minimization

- Given data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$

- Want to minimize:

$$\frac{1}{n}\sum_{i=1}^{n} \ell((\mathbf{x}_i, y_i), \mathbf{w}) + \lambda||\mathbf{w}||_2^2 = \frac{1}{n}\sum_{i}(1 - y_i((\mathbf{w}^T\mathbf{x}_i + b))_+ + \lambda||\mathbf{w}||_2^2$$

- As we've discussed, subgradient descent works like gradient descent:
  - But if there are multiple subgradients at a point, just pick (any) one:

$$\begin{aligned}
\mathbf{w}_{t+1} &:= \mathbf{w}_t - \eta\left(\frac{1}{n}\sum_{i=1}^{n} \partial_{\mathbf{w}}\ell((\mathbf{x}_i, y_i), \mathbf{w}) + 2\lambda\mathbf{w}_t\right) \\
&= \mathbf{w}_t - \eta\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{y(\mathbf{w}_t \cdot \mathbf{x}_i + b) \leq 1\}(-y_i\mathbf{x}_i) + 2\lambda\mathbf{w}_t\right) \\
&= \mathbf{w}_t + \eta\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{y(\mathbf{w}_t \cdot \mathbf{x}_i + b) \leq 1\}(y_i\mathbf{x}_i) - \eta 2\lambda\mathbf{w}_t.
\end{aligned}$$

$$\partial_{\mathbf{w}}\ell((\mathbf{x}, y), \mathbf{w}) = \mathbb{I}\{y(\mathbf{w}^T\mathbf{x} + b) \leq 1\}(-y\mathbf{x})$$

SVM

- **(Sub)gradient Descent Update**

$$
\begin{aligned}
\mathbf{w}_{t+1} &:= \mathbf{w}_t - \eta\left(\frac{1}{n}\sum_{i=1}^{n}\partial_{\mathbf{w}}\ell((\mathbf{x}_i, y_i), \mathbf{w}) + 2\lambda\mathbf{w}_t\right) \\
&= \mathbf{w}_t - \eta\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{y(\mathbf{w}_t \cdot \mathbf{x}_i + b) \leq 1\}(-y_i\mathbf{x}_i) + 2\lambda\mathbf{w}_t\right) \\
&= \mathbf{w}_t + \eta\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{y(\mathbf{w}_t \cdot \mathbf{x}_i + b) \leq 1\}(y_i\mathbf{x}_i) - \eta 2\lambda\mathbf{w}_t.
\end{aligned}
$$

- **SGD update**

$$
\mathbf{w}_{t+1} := \mathbf{w}_t + \eta\,\mathbb{I}\{y(\mathbf{w}_t \cdot \mathbf{x}_i + b) \leq 1\}(y_i\mathbf{x}_i) - \eta 2\lambda\mathbf{w}_t.
$$

Machine learning problems

- Given i.i.d. data set:

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$$

- Find parameters **w** to minimize average loss

(or regularized version):

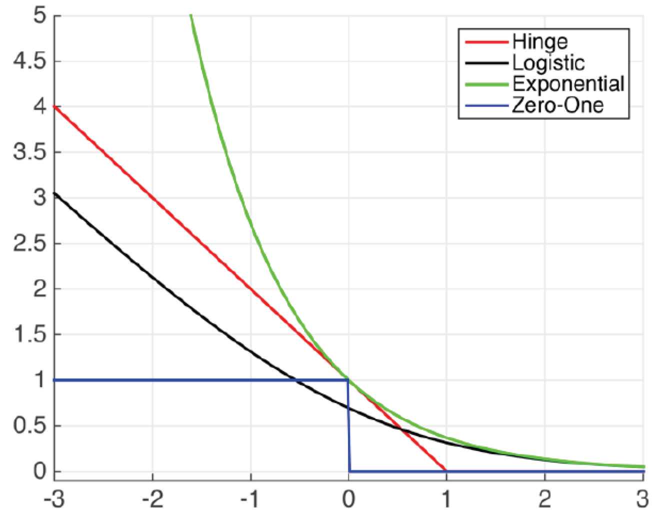
$$\frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w})$$

Squared loss: $\ell_i(\mathbf{w}) = (y_i - x_i^T \mathbf{w})^2$

Logistic loss: $\ell_i(\mathbf{w}) = \log(1 + \exp(-y_i \mathbf{x}_i^T \mathbf{w}))$

Hinge loss: $\ell_i(\mathbf{w}) = \max\{0, 1 - y_i \mathbf{x}_i^T \mathbf{w}\}$

Courtesy Killian Weinberger



What you need to know…

- Maximizing margin
- Derivation of SVM formulation
- Non-linearly separable case
  - Hinge loss
  - a.k.a. adding slack variables
- Can optimize SVMs with SGD
  - Many other approaches possible