CSE 575
Statistical Machine Learning

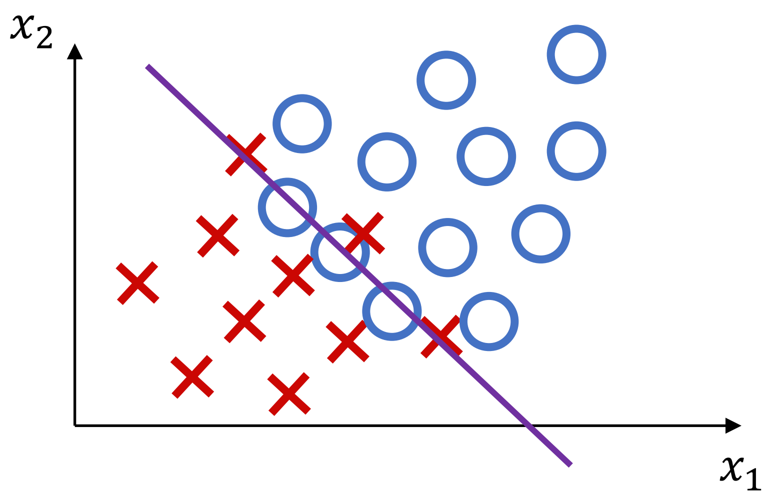
Lecture 11

YooJung Choi

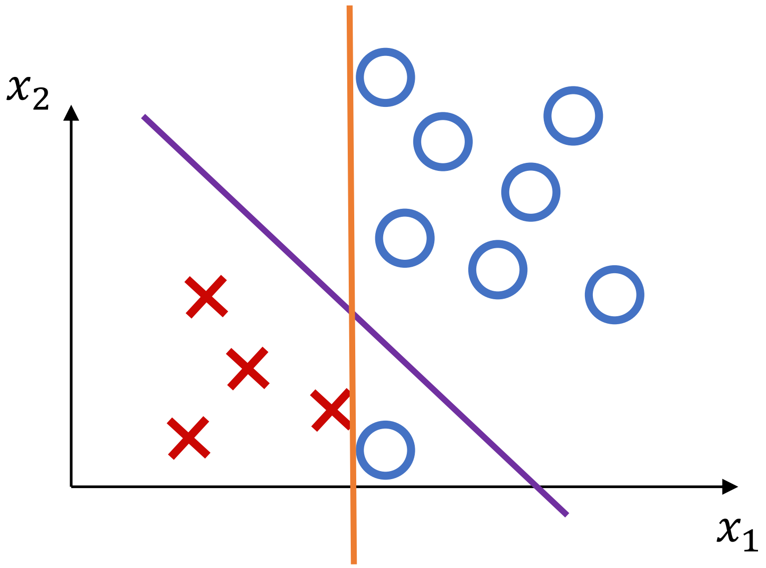
Fall 2022

Recap: Soft-margin SVM

For better generalization, we may want to allow the classifier to *make small mistakes* in exchange for *large margins*

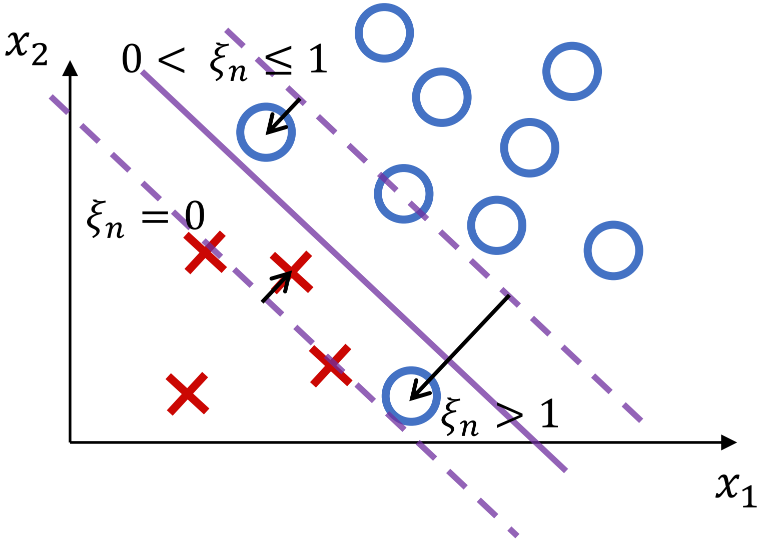


Overlapping distributions



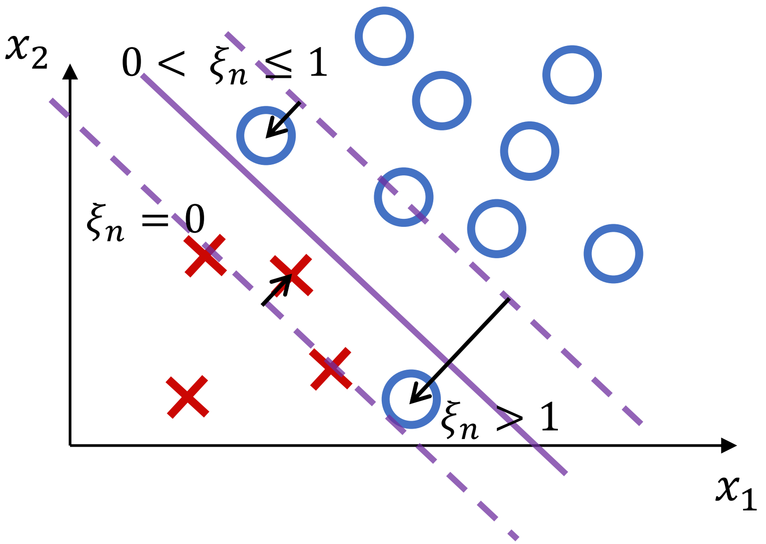
Outliers

Recap: Soft-margin SVM



- Introducing slack variables $\xi_n \geq 0$:
  - $\xi_n = 0$: $\mathbf{x}_n$ is correctly classified on the right side of margin boundary
  - $0 < \xi_n \leq 1$: $\mathbf{x}_n$ is correctly classified but on the wrong side of margin boundary (*margin violation)*
  - $\xi_n > 1$: $\mathbf{x}_n$ is misclassified
- Intuition: we want the total slack to be small

Soft-margin SVM



- For a positive example $\mathbf{x}_n$:

$$\xi_n \geq 1 - (\mathbf{w}^T \mathbf{x}_n + b)$$

- For a negative example $\mathbf{x}_n$:

$$\xi_n \geq (\mathbf{w}^T \mathbf{x}_n + b) \;-(-1)$$

- Combining using t_n:

$$\boxed{t_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n, \qquad \forall n}$$

Soft-margin SVM

Dual formulation

$$\text{argmin}_{\mathbf{w},b}\,\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{s.t.}\quad t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \quad \forall n$$

$$\xi_n \geq 0 \;\; \forall n$$

$$\text{argmax}_{\boldsymbol{a}} \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j t_i t_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.t.}\;\; 0 \leq a_n \leq C \quad \forall n$$

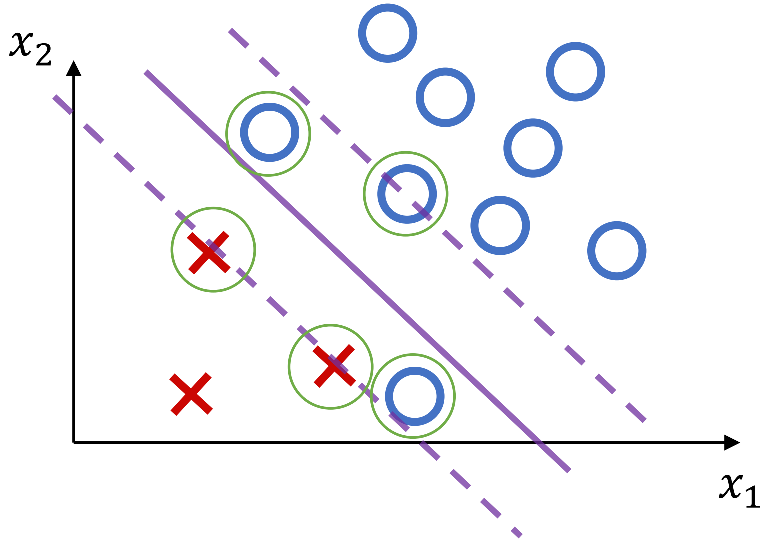
$$\sum_{n=1}^{N} a_n t_n = 0$$

- Again, *support vectors* are $\mathbf{x}_n$ such that $a_n > 0$.

  They are the examples that satisfy:

$$t_n(\mathbf{w}^T\mathbf{x}_n + b) = 1 - \xi_n$$

i.e. do not satisfy $t_n(\mathbf{w}^T\mathbf{x}_n + b) > 1$

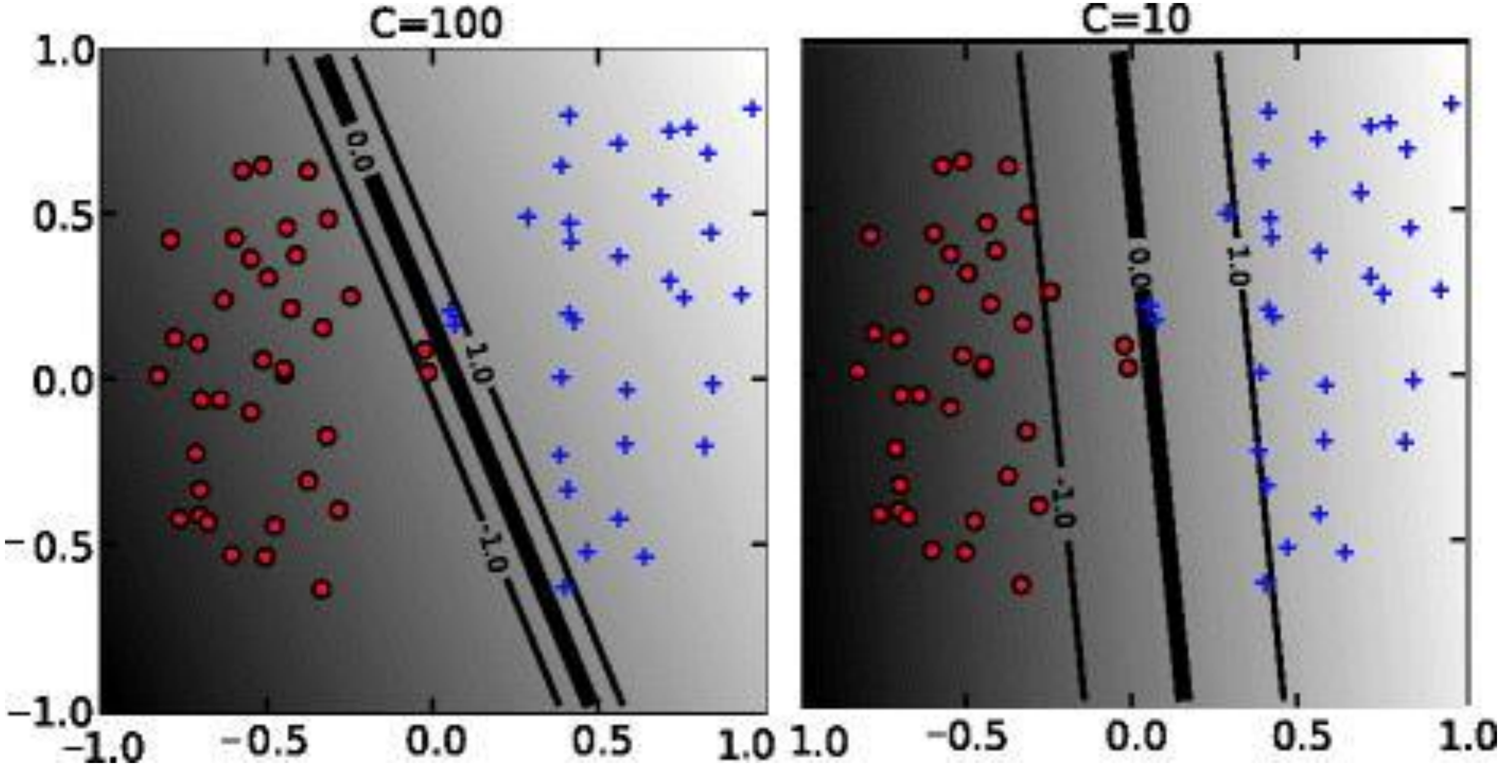


Soft-margin SVM

$$\text{argmin}_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{s.t.} \quad t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \quad \forall n$$

$$\xi_n \geq 0 \;\; \forall n$$



Hinge loss

$$\mathrm{argmin}_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{s.t.} \quad t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \quad \forall n$$

$$\xi_n \geq 0 \ \ \forall n$$

$$\mathrm{argmin}_{\mathbf{w},b} \frac{1}{n}\sum_{n=1}^{N}\xi_n + \lambda\|\mathbf{w}\|^2$$

$$\text{s.t.} \ \xi_n \geq 1 - t_n(\mathbf{w}^T\mathbf{x}_n + b) \ \ \forall n$$

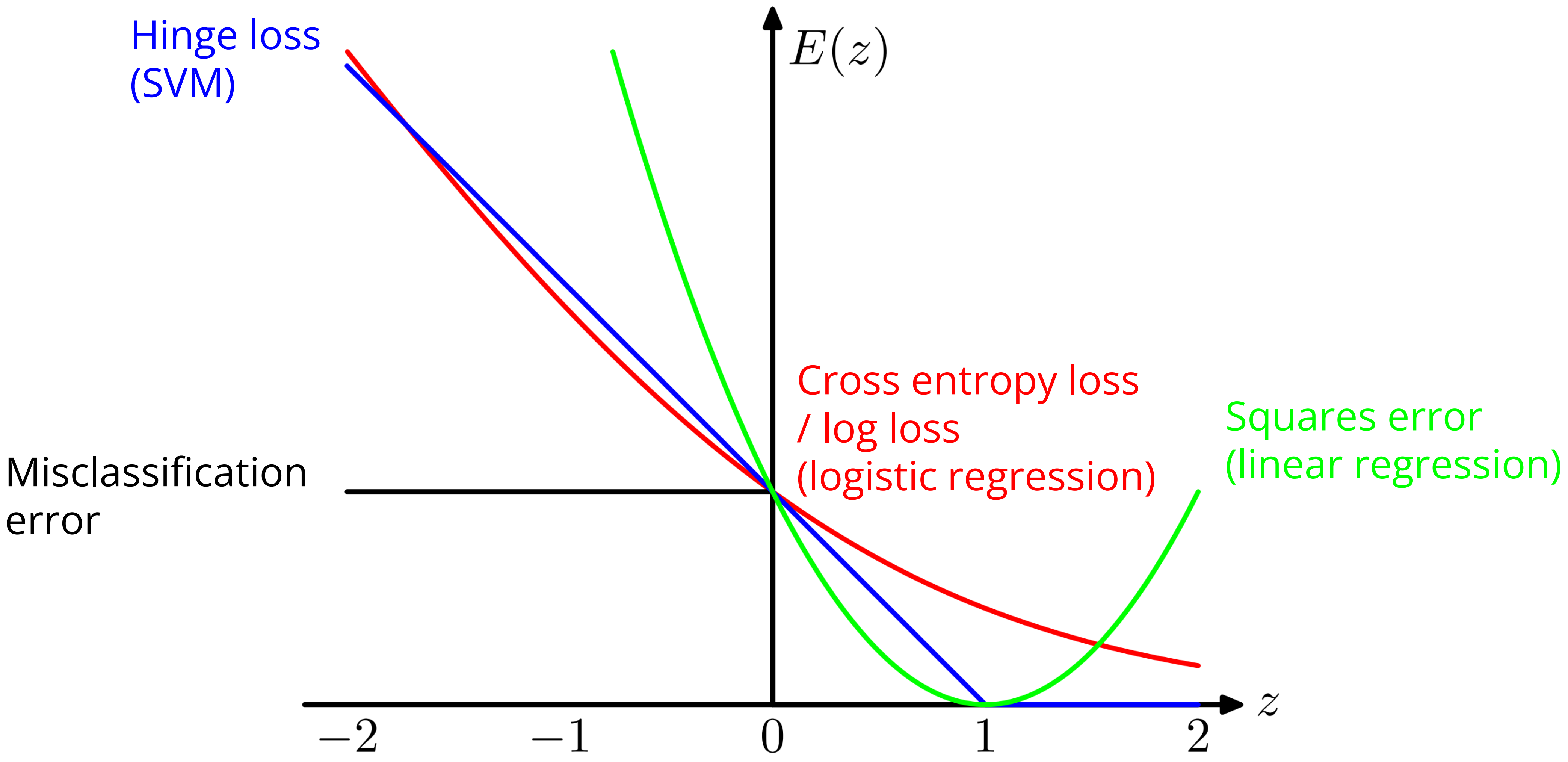
$$\xi_n \geq 0 \ \ \forall n$$

$$\mathrm{argmin}_{\mathbf{w},b} \frac{1}{n}\sum_{n=1}^{N}\xi_n + \lambda\|\mathbf{w}\|^2 \quad \text{s.t.} \ \xi_n = \max(0, 1 - t_n(\mathbf{w}^T\mathbf{x}_n + b)) \ \ \forall n$$

$$\mathrm{argmin}_{\mathbf{w},b} \underbrace{\frac{1}{n}\sum_{n=1}^{N}\max(0, 1 - t_n(\mathbf{w}^T\mathbf{x}_n + b))}_{\textit{Hinge loss}} + \lambda\|\mathbf{w}\|^2$$

L2 regularization

SVM, linear regression, logistic regression



Midterm logistics

- Written exam, Wednesday 10/5, in-class
- Closed book. You may bring a single-page letter-sized cheat sheet (with your name on it).
- No computers. Basic calculator allowed.
- Topics:
  - Probability basics
  - MLE vs MAP parameters
  - Discrete / continuous Bayes classifier
  - Naïve Bayes classifier
  - Brief questions about regression and classification algorithms

Homework 1 solutions