Machine Learning 3:
Generalization, Unsupervised learning

Hwanjo Yu

POSTECH

http://di.postech.ac.kr/hwanjoyu

Question

What is the true objective of machine learning?

① minimize error on the training set

② minimize training error with regularization

③ minimize error on unseen future examples

④ learn about machines

Roadmap

Generalization

Unsupervised learning

Training error

Loss function $\mathcal{J}(\mathbf{w})$ on training data $\mathcal{D}$:

$$\mathcal{J}(\mathbf{w}) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},y)\in\mathcal{D}} l(\mathbf{w}, \mathbf{x}, y)$$

- Find $\mathbf{w}$ that minimizes $\mathcal{J}(\mathbf{w})$:

$$\mathbf{w}^* = arg \min_{\mathbf{w}} \mathcal{J}(\mathbf{w})$$

Is this a good objective?

Rote learning

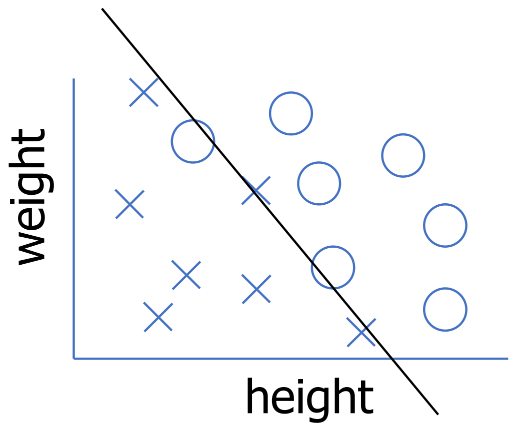
Algorithm: rote learning

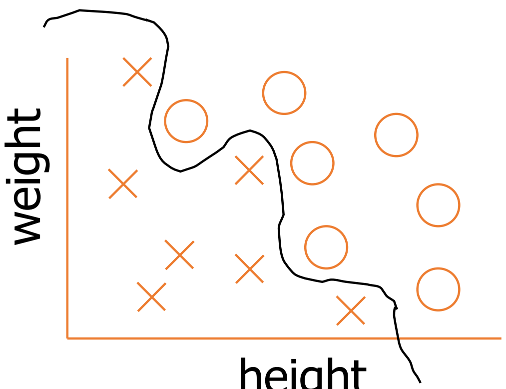
- Training: just store $\mathcal{D}$
- Predictor $f(x)$:
  If $(x, y) \in \mathcal{D}$: return y
  Else: segfault.

Minimizes the objective perfectly (zero), but clearly bad...

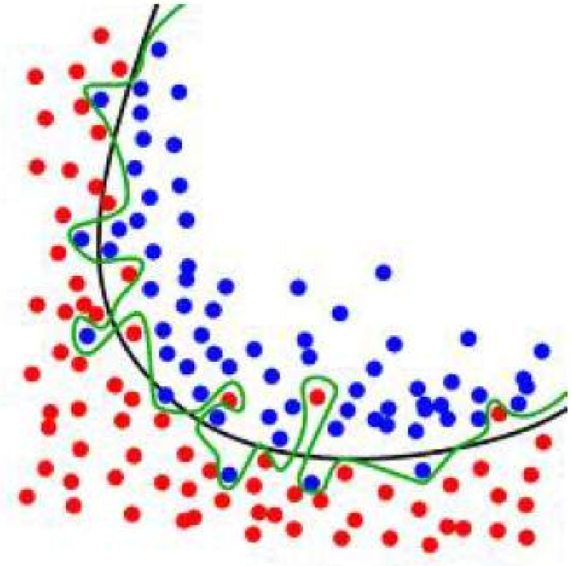
Overfitting

- x : a person; $f(x)$: male or female?
- $\phi(x) = [\text{weight}, \text{height}]$

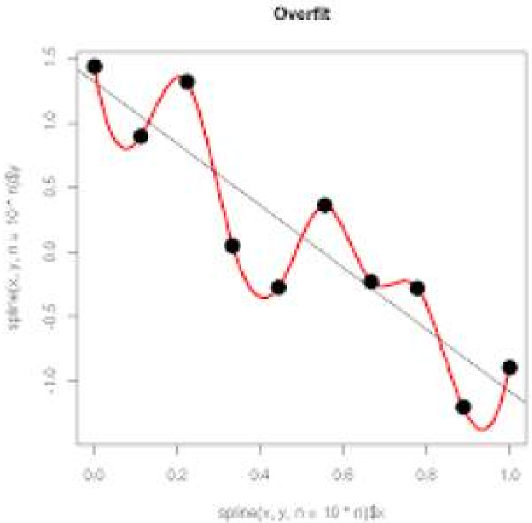




Overfitting examples

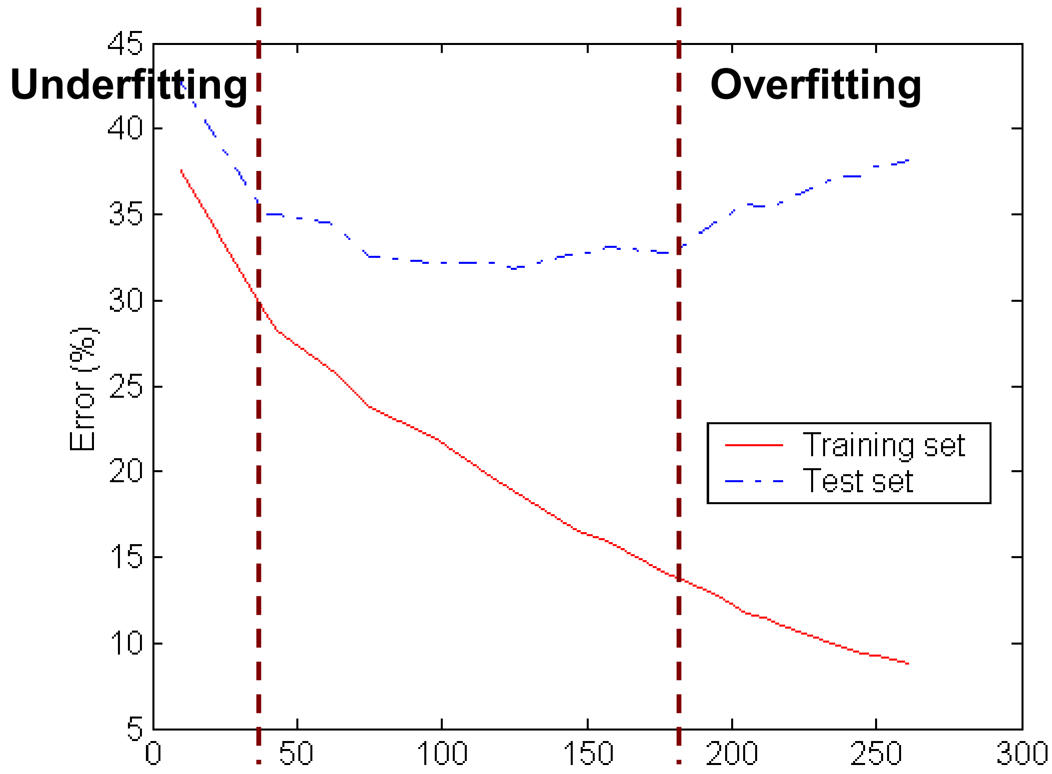


Classification



Regression

Training and test error



Model complexity (Hypothesis class)

- Underfitting: too simple
- Overfitting: too complex
- Fitting: reasonable

Question

How can you reduce overfitting (select all that apply)?

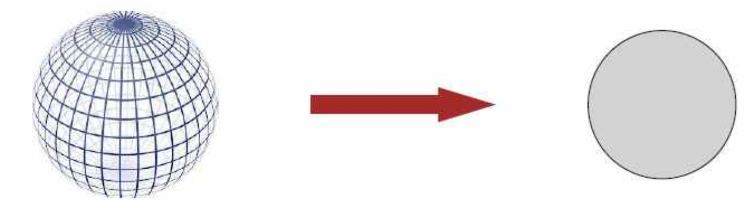
1. Remove features
2. Minimize$||\mathbf{w}||$
3. Run SGD for fewer iterations

Controlling size of hypothesis class

Example: linear predictor with weight vector $\mathbf{w} \in \mathbb{R}^d$ and $||\mathbf{w}|| \leq 1$

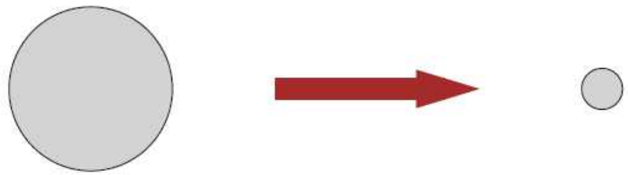
Keeping the dimensionality d small:

- $f(x) = \mathbf{w} \cdot \phi(x) = \sum_{j=1}^{d} w_j \phi(x)_j$
- $\mathbf{w} = [w_1, w_2, w_3] \Rightarrow [w_1, w_2]$



Keeping the norm (length) $||\mathbf{w}||$ small:

- $\min_{\mathbf{w} \in \mathbb{R}^d} (\text{Loss}(\mathbf{w}) + ||\mathbf{w}||)$



Controlling the dimensionality

Manual feature (template) selection:

- Add features if they help
- Remove features if they don’t help

Automatic feature selection (beyond the scope of this class):

- Wrapper method, filter method
- Forward selection, Backward selection
- Ensembles

Controlling the norm: regularization

Regularized objective:

$$\min_{\mathbf{w}} \left(\text{TrainLoss}(\mathbf{w}) + \frac{\lambda}{2} ||\mathbf{w}||^2 \right)$$

Gradient descent (GD)

- Initialize $\mathbf{w}$.
- For $t = 1, \ldots, T$:
  $\mathbf{w} \leftarrow \mathbf{w} - \alpha(\nabla_{\mathbf{w}} \text{TrainLoss}(\mathbf{w}) + \lambda \mathbf{w})$

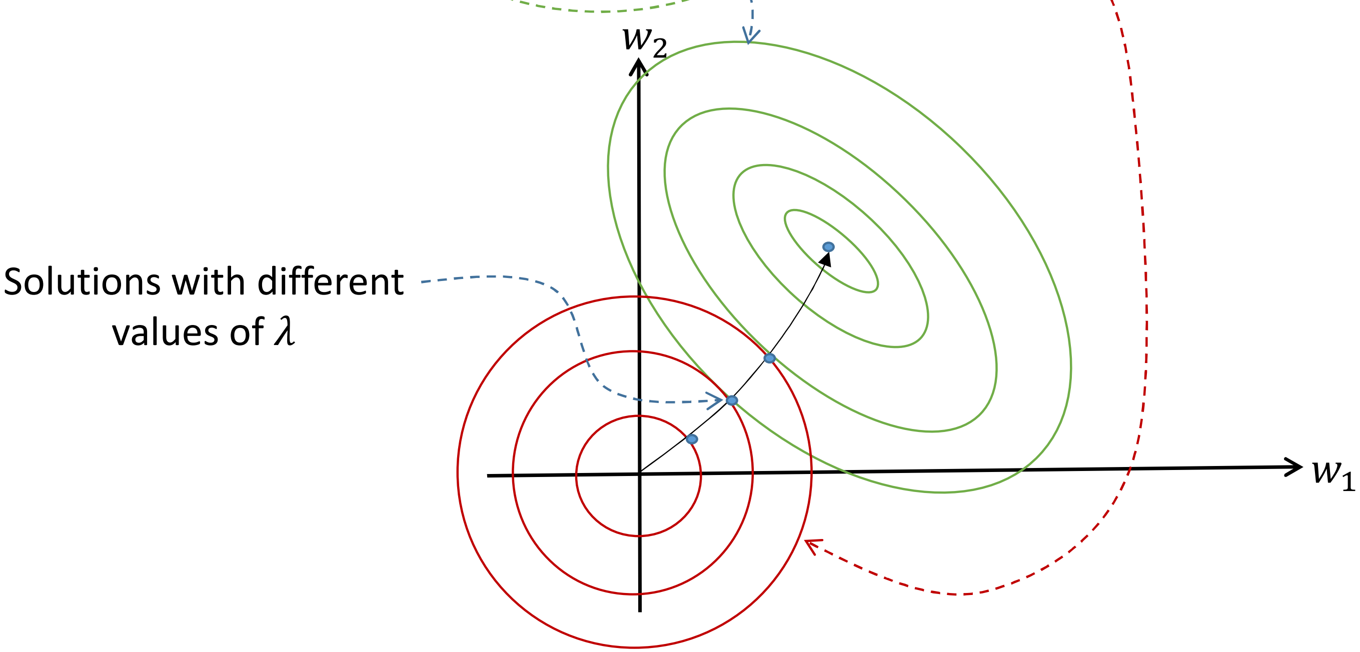
Same as gradient descent, except shrink the weights toward zero by λ

Note: SVM = hinge loss + L2 regularization

Controlling the norm: regularization

Regularized objective:

$$\min_{\mathbf{w}}(\mathrm{TrainLoss}(\mathbf{w}) + \frac{\lambda}{2}||\mathbf{w}||^2)$$



Validation

- **Validation**
  - Tuning hyperparameters: # of epoch, batch size $|\boldsymbol{\mathcal{B}}|$, step size α, regularization λ

- **Holdout test**
  - Given data is randomly partitioned into two independent sets
    - Training set (e.g. 9/10) for model construction
    - Validation set (e.g. 1/10) for accuracy estimation

What happens if we estimate accuracy on training set?

Test set

- Hyperparameters, tuned on validation set, could overfit to validation set.
- Need another set (i.e. test set) to estimate the "true" generalization error

Roadmap

Generalization

Unsupervised learning

Supervised learning vs Unsupervised learning

Supervised learning:

- $\mathcal{D}$ contains input-output pair (x, y)
- Fully-labeled data is very expensive to obtain (we can get 10,000 labeled examples)

Unsupervised learning:

- $\mathcal{D}$ only contains input x
- Unlabeled data is much cheaper to obtain (we can get 100 million unlabeled examples)

Unsupervised example: word clustering

Input: raw text (100 million words of news articles)

Output:

- Cluster 1: Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays Saturdays
- Cluster 2: June March July April January December October November September August
- Cluster 3: water gas coal liquid acid sand carbon steam shale iron
- Cluster 4: great big vast sudden mere sheer gigantic lifelong scant colossal
- Cluster 5: man woman boy girl lawyer doctor guy farmer teacher citizen
- Cluster 6: American Indian European Japanese German African Catholic Israeli Italian Arab
- Cluster 7: pressure temperature permeability density porosity stress velocity viscosity gravity tension
- Cluster 8: mother wife father son husband brother daughter sister boss uncle
- Cluster 9: machine device controller processor CPU printer spindle subsystem compiler plotter
- Cluster 10: John George James Bob Robert Paul William Jim David Mike
- Cluster 11: anyone someone anybody somebody
- Cluster 12: feet miles pounds degrees inches barrels tons acres meters bytes
- Cluster 13: director chief professor commissioner commander treasurer founder superintendent dean custodian
- Cluster 14: had hadn't hath would've could've should've must've might've
- Cluster 15: head body hands eyes voice arm seat eye hair mouth

What unsupervised learning can do?

- Data has lots of rich latent structures; want to discover this structure automatically.

- Conventional applications:
  - Density estimation
  - Clustering
  - Dimensionality reduction

- Self-supervised learning:
  - Language models
  - Generative AI
  - Anomaly detection
  - Representation learning
  - Metric learning
  - Prediction, Interpolation
  - …