جامعة قطر
QATAR UNIVERSITY

CMPS 460 – Spring 2022

MACHINE LEARNING

Tamer Elsayed

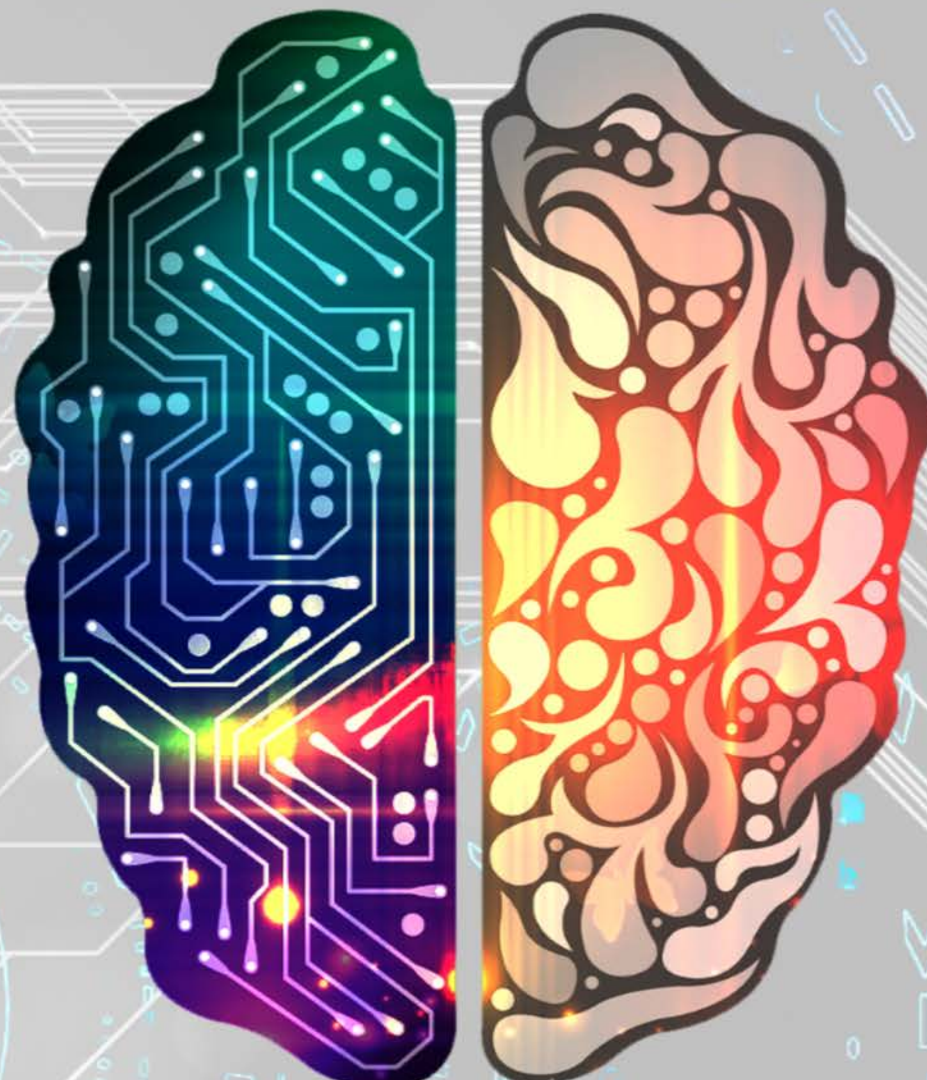


Image hosted by: WittySparks.com | Image source: Pixabay.com

7.b

Linear Models: Weight Regularization

Chapter 7: 7.3

Optimization Framework

Objective function

Loss function measures how well classifier fits training data

Regularizer prefers solutions that generalize well

$$\min_{\mathbf{w},b} L(\mathbf{w}, b) = \min_{\mathbf{w},b} \sum_{n=1}^{N} \mathbb{I}(y_n(\mathbf{w}^T \cancel{\mathbf{x}_n} + b) < 0) + \lambda R(\mathbf{w}, b)$$

$$l(y_n, \hat{y}_n)$$

λ: parameter that controls the importance of the regularization term

- Different loss function approximations
  - easier to optimize
- Regularizer
  - prevents overfitting/prefers simple models.

Surrogate loss functions

- 0-1 Loss

$$\ell^{(0/1)}(y,\hat{y}) = \mathbf{1}[y\hat{y} \leq 0]$$

- Hinge Loss

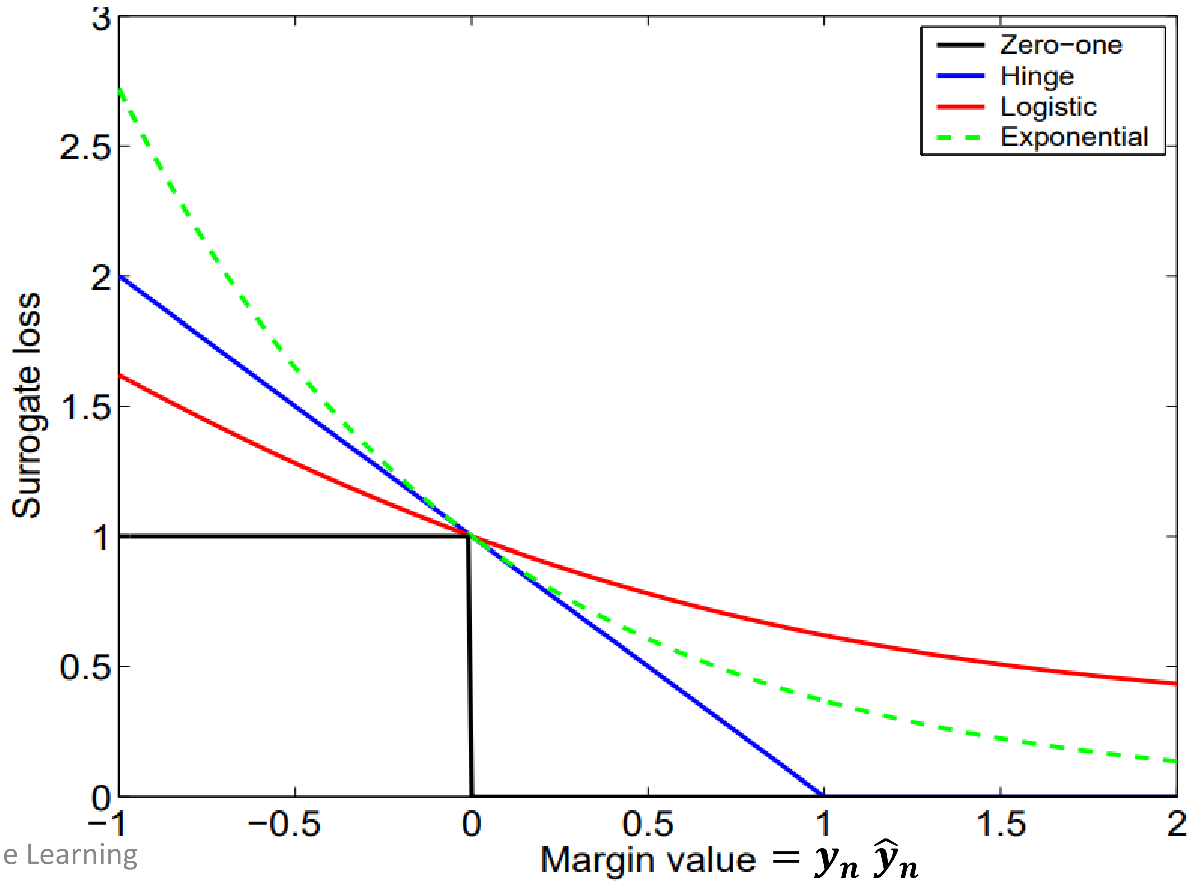
$$\ell^{(\text{hin})}(y,\hat{y}) = \max\{0, 1 - y\hat{y}\}$$

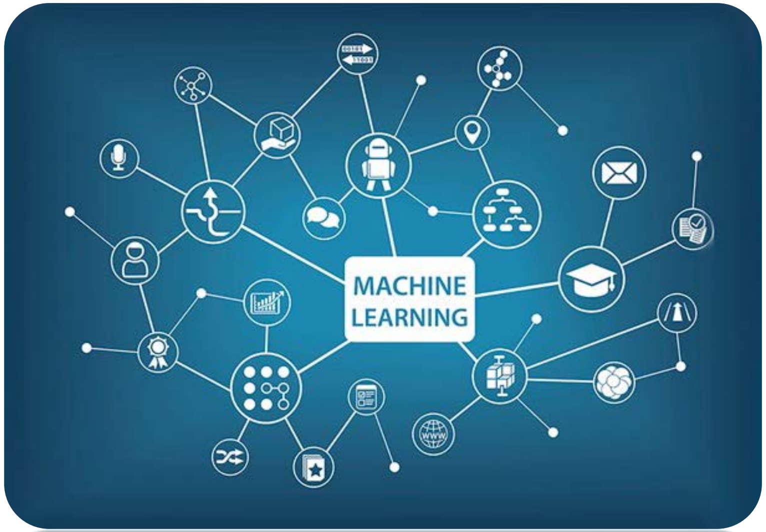
- Logistic Loss

$$\ell^{(\log)}(y,\hat{y}) = \frac{1}{\log 2}\log\left(1 + \exp[-y\hat{y}]\right)$$

- Exponential loss

$$\ell^{(\exp)}(y,\hat{y}) = \exp[-y\hat{y}]$$

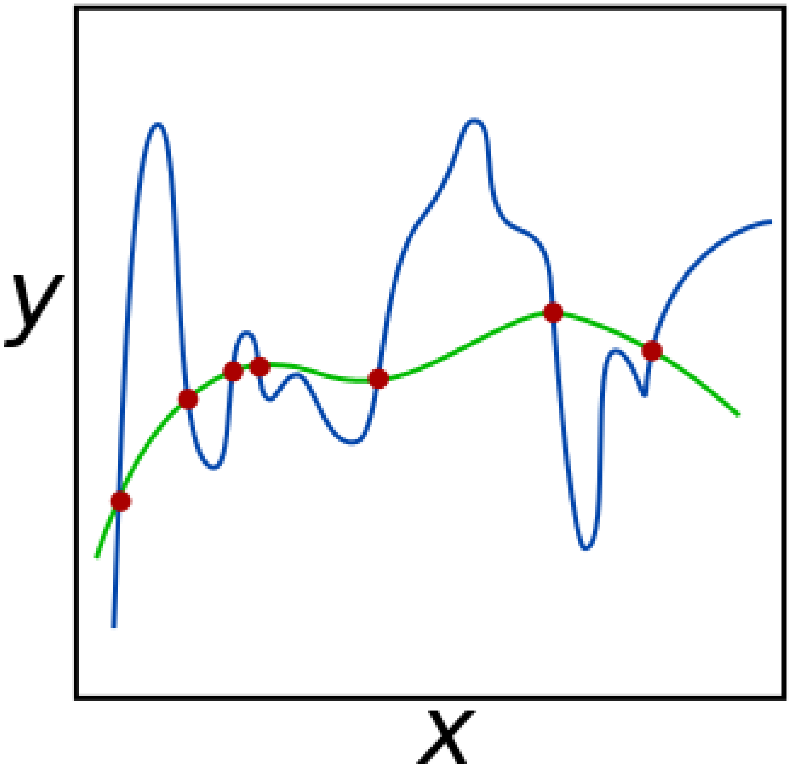




Weight Regularization

Regularization

- A technique to improve the generalizability of a learned model.
- Without bounds on complexity of the function space, model tends to overfit training data.
- Introduces a penalty for exploring certain regions of the function space.



The Regularizer Term

- Goal: find simple solutions

- Ideally, we want most entries of w to be zero, so prediction depends only on a small number of features.
- Formally, we want to minimize:

$$R^{cnt}(\mathbf{w}, b) = \sum_{d=1}^{D} \mathbb{I}(w_d \neq 0)$$

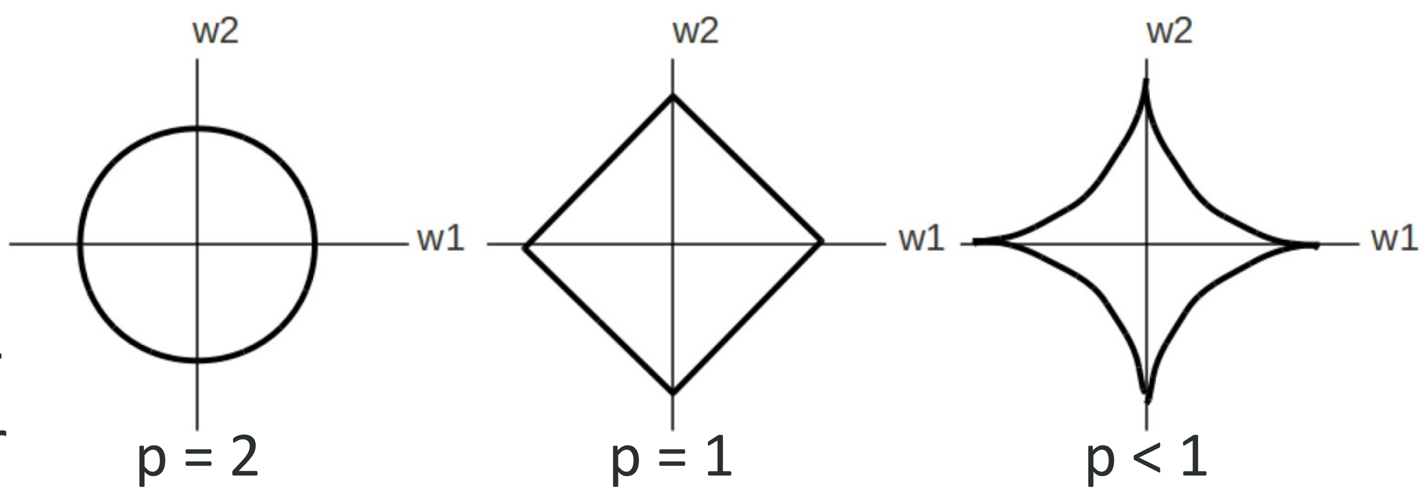
- That's NP-hard!
- So we use approximations instead.
  - e.g., we encourage w_d's to be small

Norm-based Regularizers

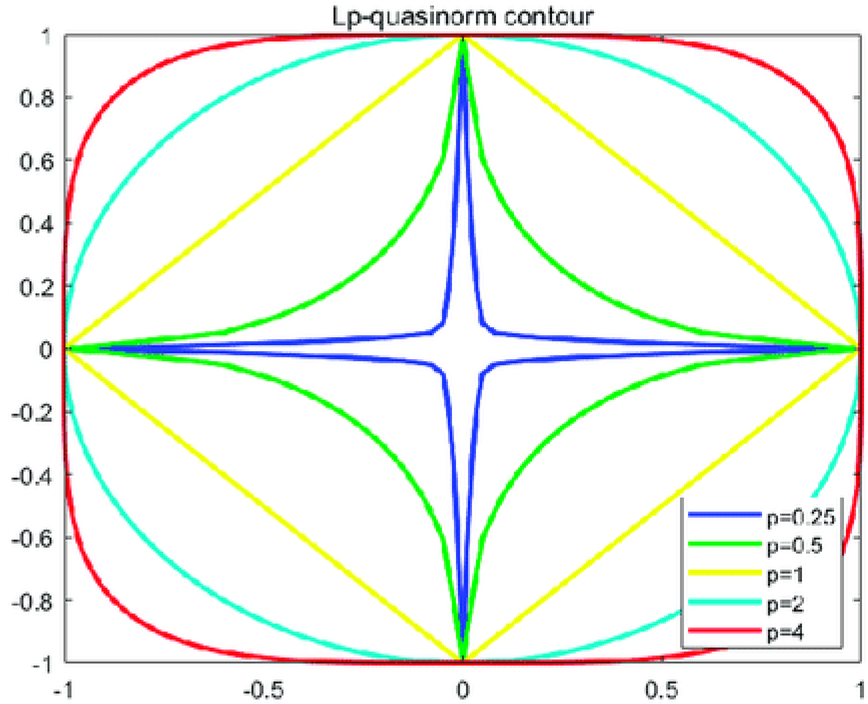
- l_p norms can be used as regularizers.

$$||\mathbf{w}||_2^2 = \sum_{d=1}^{D} w_d^2$$
$$||\mathbf{w}||_1 = \sum_{d=1}^{D} |w_d|$$
$$||\mathbf{w}||_p = (\sum_{d=1}^{D} w_d^p)^{1/p}$$

Contour plots for



Norm-based Regularizers



(a)

https://www.researchgate.net/publication/331855021_An_Efficient_Image_Reconstruction_Framework_Using_Total_Variation_Regularization_with_Lp-Quasinorm_and_Group_Gradient_Sparsity/figures?lo=1

Norm-based Regularizers

- l_p norms can be used as regularizers.

$$||\mathbf{w}||_2^2 = \sum_{d=1}^{D} w_d^2$$
$$||\mathbf{w}||_1 = \sum_{d=1}^{D} |w_d|$$
$$||\mathbf{w}||_p = (\sum_{d=1}^{D} w_d^p)^{1/p}$$

- Smaller p favors sparse vectors w
  - i.e. most entries of w are close or equal to 0
- $p < 1$: norm is non convex and hard to optimize!
- l_1 norm: encourages sparse w, convex, but not smooth at axis points
- l_2 norm: convex, smooth, easy to optimize