Exam1

Wednesday, February 17, 2021 9:17 PM

Problem # 1 :

[1] Two classes each has $\frac{n}{2}$ points overlapped

n is even.

x_1, y_1
x_2, y_2
x_3, y_3
x_4, y_4

classify x_1

$d(x_1, x_1) = 0$
$d(x_1, x_2) = d_2$
$d(x_1, x_3) = d_3$
$d(x_1, x_n) = d_n$

eg: $n = 4$, $K = 4$

$n = 6$

$K = 6$

overlap $= \frac{n}{2}$

[3]

x_1, y_1
x_2, y_2
x_3, y_3
x_4, y_4

What happen to the training error when varies

With 1 NN, the each point is itself the neighbor and the traing error is zero. As K increase to n, the error increase to $\frac{1}{2}$.

[2] K-fold cross-validation

The tradeoffs of large or small [K]

With a higher # of folds, the average estimated error will be small.

$$E = \frac{1}{K} \sum_{i=}^{K} E_i(w)$$

For a large K, the model will have low bias and high variance which leads to small average error.

[3] $$\sigma(x) = \frac{1}{1 + \exp(-w^T x)}$$

$$L_2 = \frac{\lambda}{2} \|w\|_2^2$$

Cross entropy : $$J(w) = -\sum_{i=1}^{m} \left[y_i \log \sigma(x_i) + (1 - y_i) \log(1 - \sigma(x_i)) \right]$$

$$L(w) = \min_{w} \left\{ -\sum_{i=1}^{m} \left[y_i \log \sigma(x_i) + (1-y_i) \log(1-\sigma(x_i)) \right] + \frac{\lambda}{2} \|w\|_2^2 \right\}$$

We know that $\frac{\partial J}{\partial w} = \sum_{i=1}^{m} (\sigma(x_i) - y_i)\, x_i^j$

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{m} (\sigma(x_i) - y_i) x_i^j + \lambda w$$

since this function does not have closed form solution we use gradient descent to minimize it.

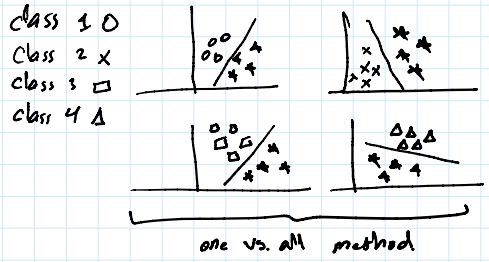
The cross entropy is a convex function.

L_2 norm is a convex function, so if it is added to a convex function ($J(w)$), the combined function will be convex.

problem #2

[1] This is because the model is overfitted to the training dataset. Adding regularization to the model and using K-fold cross validation could help solving that.

[2] We need to train our SVM 4 times.



[3] The naive Bayes approach assumes that the features of a class are conditionally independent. Meaning the presence of one feature won't affect the others

[4] We can use one of these approaches:

- Wilcoxon signed rank test to compare two classifier over multiple dataset
- Friedmen test for K classifiers over multiple dataset.

Steps

Prepare the data:

5 steps

prepare the data:

find classifiers errors for each data set

- We define the hypotheses
- set $\alpha = 0.05$
- use scipy function friedmanchisquare( ) to find P-value
  or
  wilcoxon( )
- if $P \leq \alpha$:
  Reject
  else:
  fail to Reject

problem # 3

Spring 2021 **Dept. of ECE** University of Arizona

Problem #3 – A Gambler's Ruin (10 Points)

[True/False] (1 point): The support vectors in the context of an SVM are the $\mathbf{x}_i \in \mathcal{D}_{\text{train}}$ (i.e., data set) that correspond to $\alpha_i = 0$.

$$\max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\mathsf{T}} \mathbf{x}_j \right\}$$
$$\text{s.t.} \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \text{ and } 0 \leq \alpha_i \leq C$$

The non-zero α_i are know as support vectors.

[True/False] (1 point): Increasing the term C in a support vector machine will decrease the number of support vectors.

$C \propto \frac{1}{\#\text{ of support vectors}}$

[True/False] (1 point): Parzen windows are one way to estimate the density, $p_n(x)$, but they cannot be used in classification.

[True/False] (1 point): Regularization is one way to prevent overfitting and the reason it is so effective is because the regularization term is data-independent. Therefore, the optimization process will "find" the best way to be resilient against overfitting.

[True/False] (1 point): The signed-rank Wilcoxon test is the best way to show that multiple classifiers are performing better than each other.

Wilcoxon is for one vs. one over multiple data.

[True/False] (1 point): The Friedman test is the best way to show that multiple classifiers are performing better than each other.

[True/False] (1 point): Parzen windows estimate the density of a dataset by growing a volume V until there are k samples that are enclosed on the region $\mathcal{R}$ with volume V.

Parzen window shrinks an initial region by having volume (V_n) as function of n

[True/False] (1 point): The testing error of a 1-NN classifier is always zero.

a sample x may not be in the training set.

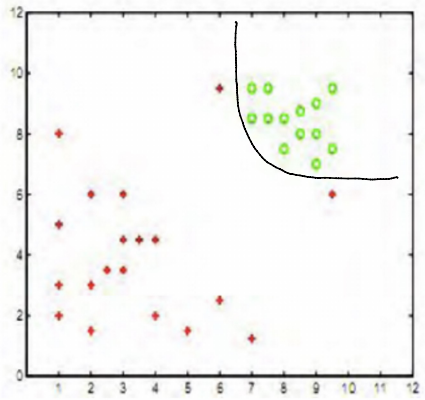
[Accept/Reject] (1 point): "My algorithm is better than yours. Look at the test error rates! (Footnote: reported results for best value of λ, chosen with 10-fold cross validation.)"

[Accept/Reject] (1 point): We did not standardize the features of our data and we did 10-fold cross validation with a k-NN classifier.

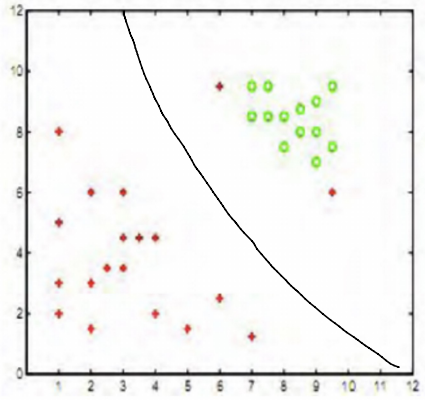
K-NN is sensitive to feature scaling.

Problem #4

[1] When C is large, the penalty is large and we have small margin. Also, the impact of slack variable is large which force the constrians to be met. (No misclassification)



2 When C is nearly equal to 0, the penalty for misclassification is small, and the margin will be maximized.



3 That phenomenon is Overfitting.

It happens when using complex model to fit training data.

problem # 5

We know that (ξ: zeda)

$$L(w, \xi, \rho, \alpha, \beta) = \frac{1}{2}\|w\|_2^2 + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i - \rho - \sum_{i=1}^{n}\alpha_i(w^T x_i - \rho + \xi_i) - \sum_{i=1}^{n}\beta_i\xi_i$$

$$\alpha_i, \beta_i \geq 0$$

We now take the derivative with respect to w, ξ, ρ

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n}\alpha_i x_i = 0 \Rightarrow \boxed{w = \sum_{i=1}^{n}\alpha_i x_i}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{\sqrt{n}} - \alpha_i - \beta_i = 0 \Rightarrow \boxed{\frac{1}{\sqrt{n}} = \alpha_i + \beta_i}$$

$$\frac{\partial L}{\partial \beta} = -1 + \sum_{i=1}^{n} \alpha_i = 0 \Rightarrow \boxed{\sum_{i=1}^{n} \alpha_i = 1}$$

- W here is different from the SVM discussed in class because

  In class, $W = \sum_{i=1}^{n} \alpha_i y_i x_i$

  In this problem, $W = \sum_{i=1}^{n} \alpha_i x_i$

  because the given data set is unlabelled.

- Yes, the constrains on α_i are

$$\boxed{\alpha_i = \frac{1}{\sqrt{n}} - \beta_i} \qquad \boxed{\sum_{i=1}^{n} \alpha_i = 1}$$

$$\boxed{0 \le \alpha_i < \frac{1}{\sqrt{n}}}$$