Bhishan poudel

HW Assignment 8 (Due by 10:30am on Dec 7)

1 Theory (150 points)

1. [**Kernel Nearest Neighbor**, 50 points]
   The nearest-neighbour classifier 1-NN assigns a new input vector $\mathbf{x}$ to the same class as that of the nearest input vector $\mathbf{x}_n$ from the training set, where in the simplest case, the distance is defined by the Euclidean metric $\|\mathbf{x} - \mathbf{x}_n\|^2$. By expressing this rule in terms of scalar products and then making use of kernel substitution, formulate the nearest-neighbour classifier for a general nonlinear kernel.

2. [**Distance-Weighted Nearest Neighbor**, 50 points]
   We have seen how to use kernels to formulate a distance-weighted nearest neighbor algorithm, when the labels are binary. Formulate a kernel-based, distance-weighted nearest neighbor that works for K classes, where $K \geq 2$.

3. [**Naive Bayes**, 50 points]
   The Naive Bayes algorithm for text categorization presented in class treats all sections of a document equally, ignoring the fact that words in the title are often more important than words in the text in determining the document category. Describe how you would modify the Naive Bayes algorithm for text categorization to reflect the constraint that words in the title are K times more important than the other words in the document for deciding the category, where K is an input parameter (include pseudocode).

4. [**Logistic Regression (*)**, 50 points]
   Assume that a binary feature x_i is equal to 1 for all training examples $\mathbf{x}$ belonging to a particular class C_k, and zero otherwise (i.e. x_i perfectly separates examples from class C_k from all other examples). Show that in this case the magnitude of the ML solution for $\mathbf{w}_k$ goes to infinity, thus motivating the use of a prior over the parameters (Hint: use the fact that the gradient on slide 24 must vanish at the solution).

2 Submission

Turn in a hard copy of your homework report at the beginning of class on the due date. On this theory assignment, **clear and complete explanations and proofs of your results are as important as getting the right answer**.

HW08

Bhishan poudel

QN1 K-Nearest Neighbor (kernel based classifier)

ⓐ 1 nearest neighbor

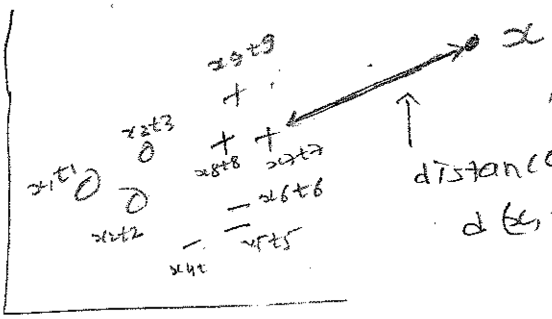
training dataset : $(x_1, t_1), (x_2, t_2), \cdots, (x_n, t_n)$

test dataset : x

required output : $\hat{y}$ (label of x or class of x)

$t_1, t_2, \cdots, t_n$ are classes from $c_1, c_2, \cdots, c_m$

for example { circle plus minus }
c1 c2 cm



(test)
predicted label = $\hat{y}$
true label = +

$$d(x, x') = \|x - x'\| = \left| \sqrt{\|x - x'\|^2} \right|$$

(Euclidean distance)

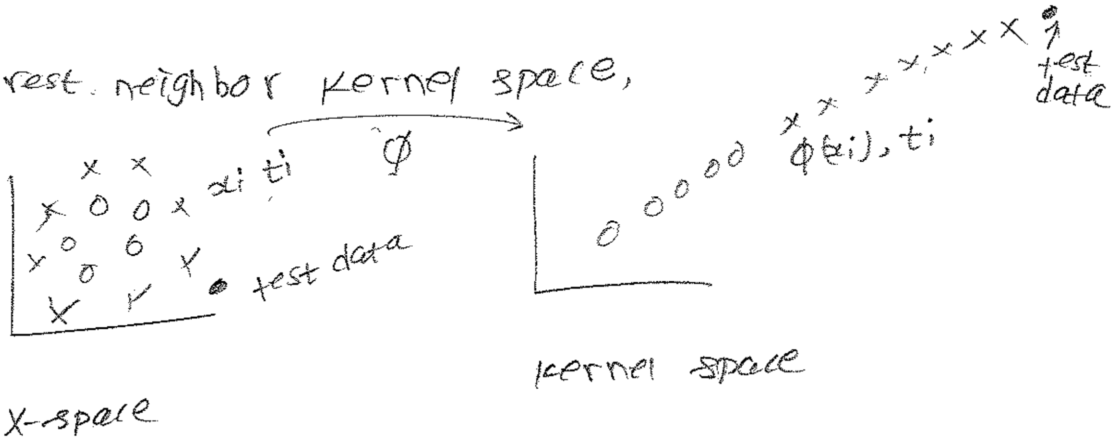
How to find $\hat{y}$?

⇒ For 1-nearest neighbor, calculate the distance (maybe Euclidean, cosine, edit, etc) from test point $\vec{x}$ to all the data points x_i, $1 < i < n$ then find the minimum distance data point x'.

Assign label of x' as the predicted label to $\vec{x}$.

$\hat{y}$ = class of x'
= class of x_i such that $d(x, x_i)$ is minimum

Ø 1-nearest neighbor kernel space,



Here, when calculating distance, we use kernel method,

x-space: $d(x,x') = \|x-x'\| = \left|\sqrt{\|x-x'\|^2}\right| = \left|\sqrt{(x-x')^T (x-x')}\right|$

(Euclidean distance)

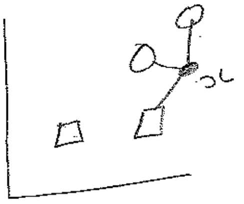
Kernel-space: $d(x,x') = \left|\sqrt{(\phi(x)-\phi(x'))^T (\phi(x)-\phi(x'))}\right|$

$= \left|\sqrt{K(x,x')}\right|$

→ we compute the distance of test $\vec{x}$ to the all the training examples x_i using kernel method then choose the minimum distance training point x' and then,

label of x = label of x'

© k-nearest neighbor kernel method.



3-nearest neighbor
two are circle
∴ label of x is circle.

Kernel ⇛ distance are calculated using kernel trick

$$d(x, x_i) = \left| \sqrt{K(x, x_i)} \right|$$

kernel can be polynomial kernel, gaussian kernel and so on.

Let $x_1, x_2, \cdots, x_k$ are nearsest of x

then,

$$\boxed{y(x) = \underset{t \in T}{\arg\max} \sum_{i=1}^{k} \delta_t(t_i)}$$

Here, $t_i = \{ 0, \square]$

$k = 3$

for x_1 and x_2 $t = t_i = 0$ number $= 2$

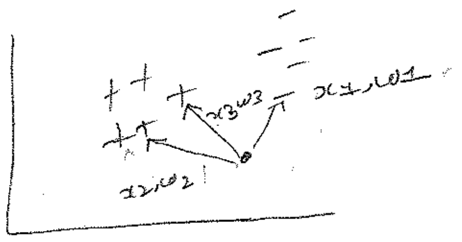
for x_3 $t = t_i = \square$ number $= 1$

argmax $= 0$

QN2 Distance weighted Nearest Neighbor

ⓐ distance-weighted KNN binary classification

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



3-nearest neighbor

contribution from + : $w_2 + w_3$

contribution from − : w_1

suppose $w_2 + w_3 > w_1$

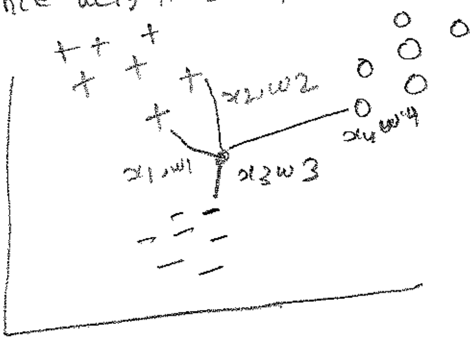
then label of x = +

- x is test example
- $x_1, x_2, \ldots, x_k$ are the nearest points to x, calculated using Euclidean distance in kernel space.
- $w_1, w_2, \ldots, w_k$ are the weights to the points $x_1, x_2, \ldots x_k$ calculated in kernel space,

$$w_i = \frac{1}{\|x - x_i\|^2} = \frac{1}{\sqrt{K(x, x_i)}}$$

- $$\boxed{\hat{y} = \text{sign} \sum_{i=1}^{K} w_i \, \delta_t(t_i)}$$

ⓑ distance weighted KNN for multiclass classification



Note: nearer points have more contribution

4-nearest neighbor

contribution from class $C_1 = +$
$= w_1 + w_2$

contribution from class $C_2 = -$
$= w_3$

contribution from cass $C_3 = 0$
$= w_4$

suppose $w_1 + w_2 > w_3$ ∴ $\hat{y} = +$
$w_1 + w_2 > w_4$

$$w_i = \frac{1}{\|x - x_i\|^2} = \frac{1}{\sqrt{K(x, x_i)}}$$

$$\boxed{\hat{y} = \underset{t \in T}{\text{argmax}} \sum_{i=1}^{k} w_i \, \delta_t(t_i)}$$

Ans

$\boxed{\text{ON3}}$ Naïve Bias

@ simple text ccassification using Naive Bias

- each document is an example vector
- a document x has n words $\omega_1, \omega_2, \ldots, \omega_n$
- from all the documents $D = \{x_1, x_2, \ldots, x_m\}$ we create vocabulary file (this is feature vector for whole dataset D)
- vocabulary $V = \{w_1, w_2, \ldots, w_{|V|}\}$, there are $|V|$ number of words in vocabulary.
- each document x belongs to one of the catagories $C_K = \{c_1, c_2, \ldots, c_K\}$

$$\begin{array}{cccccc} & & & \text{input data} & & \text{target/class/label} \\ \text{for example} & D = & \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = & \begin{bmatrix} \text{line one} \\ \text{line two} \\ \text{line three} \end{bmatrix} & \begin{matrix} \rightarrow \\ \rightarrow \\ \rightarrow \end{matrix} & \begin{matrix} \text{news} \\ \text{sports} \\ \text{religion} \end{matrix} \end{array}$$

we want ⇒ given a document x find its catagory
$p(C_K|x)$ probability of document x belonging to catagory C_K (eg news)

Naive Bias gives ⇒
① $P(C_K)$ probability of each categories (priors)
② $P(w_i|C_K)$ conditional probablility of data x w.r.t category C_K.

then we use Baye's theorem to get $p(c^*|x)$.

$$p(c^*|x) = \frac{p(x|c^*)\,p(c^*)}{p(x)}$$

$$= p(c^*) \prod_{i=1}^{n} p(x_i|c^*)$$

→ where c^* corresponds to the $p(w_i|C_K)$ that has maximum probability.

$$c^* = \underset{C_K}{\operatorname{argmax}}\; p(C_K|x)$$

The preucode when all words have equal importance,

```
for each category $C_K$:
    let $D_K$ be the subset of documents in category $C_K$
    set $P(C_K) = |D_K| / |D|$
    let $n_K$ be total numer of words in $D_K$
    for each word $w_i \in V$:
        let $n_{Ki}$ = # of $w_i$ in $D_K$
        set $P(w_i|C_K) = \frac{n_{Ki}+1}{n_K+|V|}$   ← Laplace smoothing
```

from this subroutine, we get $\to P(C_K)$

$\to P(w_i|C_K)$

then we find the posterior probability,

$$P(C_K|x) = \frac{P(x|C_K)\cdot P(C_K)}{P(x)}$$

$$\propto P(x|C_K)\, P(C_K)$$

$$\propto P(C_K)\cdot \prod_i (w_i|C_K)$$

↑ w_i are words in document x

Example

Train	DOC	words	class	Total
	1	chinese Beiging chinese	$c_1=c$	$n_1=8$ words $d_1=3$ examples
	2	chinese chinese shanghai	c_1	
	3	chinese macao	c_1	
	4	Tokyo Japan chinese	c_2	$n_2=3$ words $d_2=1$ example
test	1	chinese Chinese chinese Tokyo Japan $(\vec{x})$	?	11 total words 6 unique words

$|V|=6$

vocabulary :	{ chinese,	Beiging,	shanghai,	macao,	Tokyo,	Japan }
wordfreq in c1 :	5	1	1	1	0	0
word freq in c2 :	1	0	0	0	1	1
Laplace smoothing :	1	1	1	1	1	1

$|V|=6$

② category $c_1=c$

- prior $P(c_1) = |D_1| / |D| = 3/4$ (3 documents belongs to class c1)
- number of words in class, $n_1 = 8$
- conditional probabilities $\left(P(w_i|c_k) = \frac{n_{ki}+1}{n_k+|V|}\right)$

chinese $P(w_1|c_1) = \frac{\overset{\nearrow n_{ki}}{5} + \overset{\nearrow \text{Laplace}}{1}}{8+6} = \frac{6}{14} = \frac{3}{7}$

Beiging $P(w_2|c_1) = \frac{1+1}{14} = \frac{2}{7}$

shanghai $P(w_3|c_1) = \frac{2}{7}$

macao $P(w_4|c_1) = \frac{2}{7}$

Tokyo $P(w_5|c_1) = \frac{0+1}{8+6} = \frac{1}{14}$

Japan $P(w_6|c_1) = \frac{0+1}{14} = \frac{1}{14}$

category $c_2=j$

- prior $P(c_2) = |D_2| / |D| = 1/4$ (1 document belongs to class c2)
- $n_2 = 3$
- conditional probabilities

$P(w_1|c_2) = \frac{1+1}{3+6} = \frac{2}{9}$

$P(w_2|c_2) = \frac{0+1}{3+6} = \frac{1}{9}$

$P(w_3|c_2) = \frac{1}{9}$

$P(w_4|c_2) = \frac{1}{9}$

$P(w_5|c_2) = \frac{2}{9}$

$P(w_6|c_2) = \frac{2}{9}$

③ Choosing class

$$p(c_1|x) \propto p(c_1) \prod_i p(w_i|c_1) = \overset{\text{prior}}{\frac{3}{4}} \cdot \overset{\text{chinese}}{\left(\frac{3}{7}\right)^3} \cdot \overset{\text{Tokyo}}{\frac{1}{14}} \cdot \overset{\text{Japan}}{\frac{1}{14}} = 0.0003$$

$$p(c_2|x) \propto p(c_2) \prod_i p(w_i|c_2) = \frac{1}{4} \cdot \left(\frac{2}{9}\right)^3 \cdot \frac{2}{9} \cdot \frac{2}{9} = 0.0001$$

$$c^* = \underset{c_k}{\arg\max}\ P(c_k) \prod_{j=1}^{n} P(w_j|c_k) = \underset{c_k}{\arg\max} \{0.0003, 0.0001\} = \text{Label of } 0.0003 = \boxed{c_1}$$

P.T.O.

(b) give more importance to Title of doc

```
(1) for each class in ck :
      findclass prior $p(c_k) = |D_k| / |D|$ (where $D_k$ is subset of D with
                                              class $c_k$)
(2)
(3)   # loop through title words
      let nk = number of title words in class ck all examples
      for each titleword $w_i \in V1$ :
          let nki = number of words in title of all class ck examples

          # find conditional prob of each title words
          $P_1(w_i | c_k) = \frac{n_{ki}+1}{n_k+|V1|} \times K$ ← multiplication factor
                                                     for title words

(4)   # loop through body words
      let nk = number of body words
      for each body-word $w_i \in V2$ :
          let nki = number of words in body of all
                    class ck examples (for example ck is
                                       class1)

          # find conditional prob of each body words
          $P_2(w_i | c_k) = \frac{n_{ki}+1}{n_k+|V2|}$

          ⟹ output $P(c_k)$, $P_1(w_i|c_k)$, $P_2(w_i|c_k)$
```

choosing class:

we again suppose title words and body words are independent,

$$p(c_k | x) \propto \underbrace{P_1(x|c_k)p(c_k)}_{\text{for title}} \cdot \underbrace{P_2(x|c_k)P(c_k)}_{\text{for body}}$$