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

Bishan Poudel

Q.N.1 K-Nearest Neighbor (kernel based classifier)

(a) 1 nearest neighbor

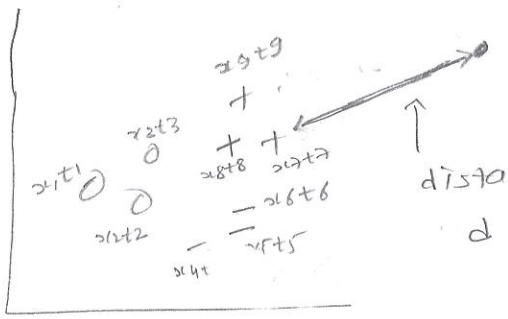
training dataset : $(x_1, t_1), (x_2, t_2), \cdots, (x_n, t_n)$

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 }
c_1 c_2 c_m



How to find $\hat{y}$?

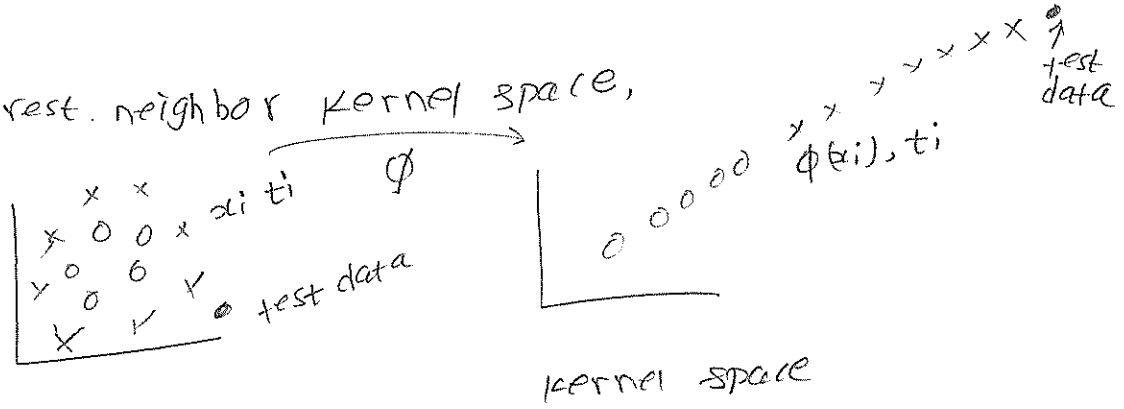
$\Rightarrow$ For 1-nearest neighbor, calculate the distance (maybe Euclidean, cosine, edit, etc) from test point $\vec{x}$ to all the data points x_i, $1 < i < n$ then find the minimum distance data point x'.

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

① 1-nearest neighbor kernel space,



X-space

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)

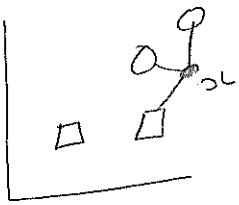
Kernel-space: $d(x, x') = \left|\sqrt{(\phi(x) - \phi(x'))^T (\phi(x) - \phi(x'))}\right|$

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

→ 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'

⑥ K-nearest neighbor kernel method.



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

Kernel ⇒ distances are calculated using kernel trick

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

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

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

then,

$$y(x) = \underset{t \in T}{\operatorname{argmax}} \sum_{i=1}^{K} \delta_t(t_i)$$

Here, $t_i = \{$ circle, square $\}$

$K = 3$

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

for x_3 $t = t_i =$ □ number $= 1$

argmax $= ○$

summary

input $(x_1, t_1), (x_2, t_2), \cdots (x_n, t_n)$ $t_n \in \{c_1, c_2, \cdots c_K\}$

test x, $t = ?$

map input from x-space to kernel-space

x-space input $\phi(x_1)$ $\phi(x_2)$ $\phi(x_n)$

kernel space test $\phi(x)$, $t = ?$

compute similarity $k(x, x_i)$ for all x_i, $1 \le i \le n$

(compute distance $= \sqrt{k(x, x')}$ is redundant since distance $\propto$ similarity)

then predicted class of x is,

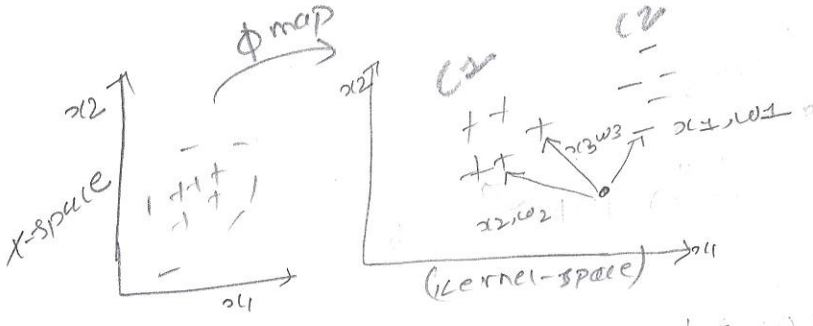
$$y(x) = \underset{t \in T}{\operatorname{argmax}} \sum_{i=1}^{k} \delta_t(t_i)$$

Q2 Distance-weighted Nearest Neighbor for classification

kernel space

(a) distance-weighted KNN binary classification

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



3-nearest neighbor

contribution from +: $w_2 + w_3$

contribution from −: w_1

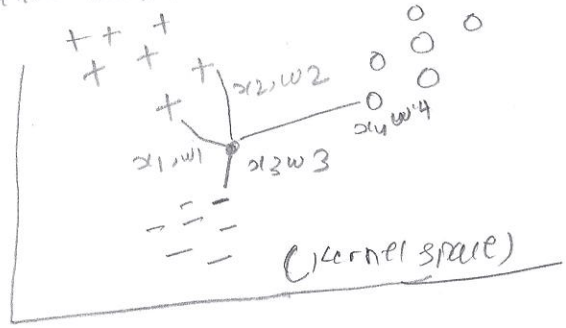
suppose $w_2 + w_3 > w_1$

then label of $x = +$

- x is test example
- $x_1, x_2, \dots, x_k$ are the nearest points to x, calculated using Euclidean distance in kernel space.
- $K(x, x_i)$ is the measure of similarity between point x and x_i in kernel space.

$$\hat{y} = \text{sign} \sum_{i=1}^{k} K(x, x_i)\, \delta_t(t_i)$$

(b) kernel-based 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 class $c_3 = 0$
$= w_4$

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

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

Ans

DW3 Naïve Bias

① simple text classification using Naive Bias

→ each document is an example vector

→ a document x has n words $w_1, w_2, \dots, w_n$

→ from all the documents $D = \{x_1, x_2, \dots, x_m\}$

we create vocabulary file (this is feature vector for whole dataset D)

→ vocabulary $V = \{w_1, w_2, \dots, w_{|V|}\}$, there are $|V|$ number of words in vocabulary.

→ each document x belongs to one of the catagories $C_K = \{c_1, c_2, \dots, c_K\}$

for example $D = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \text{line one} \\ \text{line two} \\ \text{line three} \end{bmatrix}$ (input data) $\begin{bmatrix} \text{news} \\ \text{sports} \\ \text{religion} \end{bmatrix}$ (target (class) label)

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 probability of data x w.r.t. category c_k.

then we use Bayes 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

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

NB for text classification when all words have equal importance

(A)
- D: Documents $x_1, x_2, \ldots x_n$
- $\{c_k\}$: classes $c_1, c_2, \ldots c_K$
- V: vocabulary (unique words in D − stopwords − below threshold words) $\{w_1, w_2, \ldots, w_V\}$
- D_k: subset examples having class c_k
- n_k: # of words in D_k

(B)
1. for each category c_k:
2. D_k = subset with category c_k
3. $P(c_k) = \frac{|D_k|}{|D|}$ (prior)
4. n_k = # words in D_k
5. for each word $w_i \in V$:
6. let n_{ki} = # of w_i in D_k
7. set $p(w_i|c_k) = \frac{n_{ki} + 1}{n_k + |V|}$ ← Laplace smoothing
   (cond prob)
8. Return priors $P(c_k)$ and cond prob $p(w_i|c_k)$

(C)
Then we find the posterior probability for each class c_k,

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

$$\propto p(x|c_k) P(c_k)$$

$$\propto P(c_k) \cdot \prod_{i=1}^{n_V} p(w_i|c_k)$$ (w_i are words in test document x)

$$\propto \ln \left[P(c_k) \cdot \prod_i p(w_i|c_k) \right]$$

$$\propto \ln P(c_k) + \sum_i \ln (p(w_i|c_k))$$

(to prevent underflow)

(D)
Testing

$$c^* = \arg\max_{c_k} p(c_k|x)$$

$$c^* = \arg\max_{c_k} P(c_k) \prod_i p(w_i|c_k)$$

$$\propto \arg\max_{c_k} \left[\ln P(c_k) + \sum_i \ln (p(w_i|c_k)) \right]$$

Example of Naive Bias

train	Doc	words	class	Total
	1	chinese Beiging chinese	$c_1=c$	$n_1=8$ words $d_1=3$ examples $n_1+\lvert V\rvert = 8+6=14$
	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 $n_2+\lvert V\rvert = 3+6=9$
test	1	chinese Chinese chinese Tokyo Japan ($\vec{x}$)	?	11 total words 6 unique words $\lvert V\rvert = 6$

vocabulary : { chinese (w_1), Beiging (w_2), shanghai (w_3), macao (w_4), Tokyo (w_5), Japan (w_6) } $\lvert V\rvert = 6$

	w_1	w_2	w_3	w_4	w_5	w_6
word freq in c_1 :	5	1	1	1	0	0
word freq in c_2 :	1	0	0	1	1	1
Laplace smoothing :	1	1	1	1	1	1
$\frac{n_{ki}+1}{n_k+\lvert V\rvert}$:	6, 2 14, 9	2 14, 9	1 14, 9	2 14, 9	1, 2 14, 9	1, 2 14, 9

② category $c_1=c$

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

Chinese $P(w_1 \mid c_1) = \frac{5+1}{8+6} = \frac{6}{14} = \frac{3}{7}$ (n_{ki} → number of word w_i in class c)

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

shanghai $P(w_3 \mid c_1) = \frac{1+1}{14} = \frac{2}{7}$

macao $P(w_4 \mid c_1) = \frac{1+1}{14} = \frac{2}{7}$

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

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

category $c_2=j$

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

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

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

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

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

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

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

③ choosing class

prior chinese tokyo japan

$P(c_1 \mid x) \propto P(c_1) \prod_{i \in c_1} P(w_i \mid c_1) = \frac{3}{4} \cdot \left(\frac{3}{7}\right)^3 \cdot \frac{1}{14} \cdot \frac{1}{14} = 0.0003$

$P(c_2 \mid x) \propto P(c_2) \prod_{i \in c_2} P(w_i \mid c_2) = \frac{1}{4} \cdot \left(\frac{2}{9}\right)^3 \cdot \frac{2}{9} \cdot \frac{2}{9} = 0.0001$

$c^* = \underset{c_k}{\operatorname{argmax}}\ P(c_k) \prod_{j=1}^{n} P(w_i \mid c_k) = \underset{c_k}{\operatorname{argmax}} \{0.0003, 0.0001\} =$ label of $0.0003 = c_1$

P.T.O.

⑤ give more importance to title of doc

① for each category c_k :

② bindclass prior $P(c_k) = |D_k| / |D|$ (where D_k is subset of D with class c_k)

③ # loop through title words
let n_k = number of title words in class c_k all examples
for each titleword $w_i \in V_1$:
let n_{ki} = number of words in title of all class c_k examples
. . . . # bind conditional prob of each title words
$P_1(w_i | c_k) = \frac{n_{ki}+1}{n_k + |V_1|} \times K$ ← multiplication factor for title words

④ # loop through body words
let n_k = number of body words
for each bodyword $w_i \in V_2$:
let n_{ki} = number of words in body of all class c_k examples (for example c_k is class 1)
bind conditional prob of each body words
$P_2(w_i | c_k) = \frac{n_{ki}+1}{n_k + |V_2|}$

⟹ 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,

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