CS 4780/5780 Machine Learning Due September 11, 2013

Assignment 1: Solutions: Concept Learning and kNN

Instructor: Thorsten Joachims *Total Points: 100*

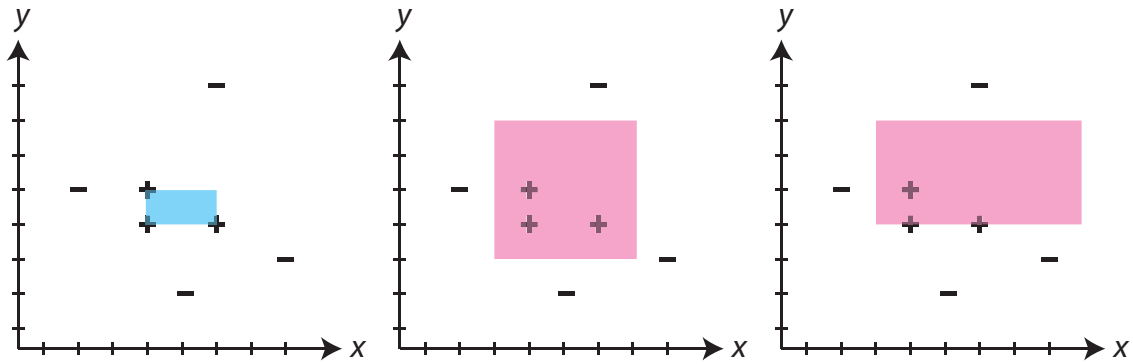
Problem 1: Version Spaces [30 points]

(a) *Rectangular Hypothesis space* The size of the hypothesis space is a little tricky to compute and involves careful counting. The number of hypotheses $|H|$ is essentially $N_{rectangles} + N_{lines} + N_{points}$. Number of rectangles is half of the number of possible combinations of its two corners (but we have to subtract out the degenerate cases first). Then we add back the degenerate line and point cases and 1 $NULL$ case to get (where $n = 9$):

$$|H| = \frac{1}{2}\binom{n^2}{2} + n\binom{n}{2} + n^2 + 1 = 2026$$

There are other ways to obtain the same answer.

(b) *Most General/Specific* There is only one most specific hypothesis and multiple most general hypotheses:



(a) Most specific H (b) Most General H (c) Most General H

Figure 1: Problem 1*b*

(c) *Version space* The top left corner can move to 6 possible points, the bottom right corner to 5, before violating any training examples, therefore $|V| = 30$

(d) *Rectangular hypothesis* Using the same approach to counting version space as in the previous part, consider each of the candidate query points as either positive or

negative. Counting should yield an expected $|V| = 15$ for all three candidate queries. You can also show that in general, for an equal probability of positive and negative class, the expected value of the version space after observing another instance (inside the current version space) is half the size of the current version space. Consider some query instance x_q inside the current version space (such as the three points given in the problem) observed with a positive label. The resulting version space size is then $|V_q^+|$, such that $1 \leq |V_q^+| \leq |V|$. If x_q turns out to be negative, then $|V_q^-| = |V| - |V_q^+|$. The expected size of the version space is then $1/2|V_q^+| + 1/2|V_q^-| = 1/2|V|$.

(e) *Decision Tree hypothesis 1-level* There is no 1-level decision tree that will separate all instances without training error. Thus, $|V| = 0$

(f) *Decision Tree hypothesis 3-leaf* There are 5 possible functions that can be produced with a 3-leaf decision tree described in the problem. Displayed below are the decision boundaries

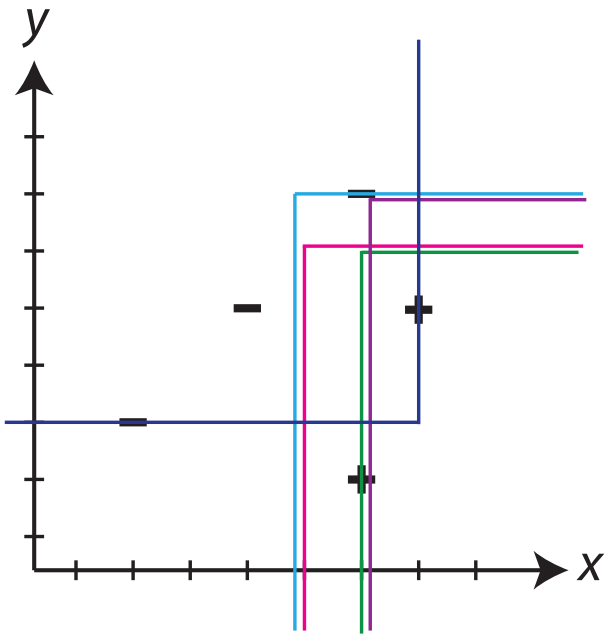


Figure 2: Problem $1f$

Problem 2: Regression with kNN [30 points]

Displayed below are the corresponding face completions for $K = 5$ and $K = 50$. The main observation is that for a larger value of K, the image becomes blurrier. Larger K averages more faces together, causing the blur.

Figure 3: Face regression K=5

Figure 4: Face regression K=50

Problem 3: kNN Classification [40 points]

(a) *Content-based book recommendation* : For the *Fifty-Shades* book, the top recommendations are *Fifty Shades Freed, Secrets Vol. 2, SHADES OF CONDEMNATION, The Edge of Never, Late, Hopless*

For the *Brains: A Zombie Memoir*, the top recommendations are *Criminal: A Novel, Dying to Live, War Against the Walking Dead: The Ministry of Zombies, Brownies, Easter Bunny Murder*

(b) *Baseline*. See Table 1 Accuracy = 62.2%

class	Precision	Recall
0	0.53	0.55
1	0.43	0.44
2	0.53	0.61
3	0.73	0.67
4	0.65	0.58

Table 1: Centroid baseline part *b*

(c) *kNN Implementation* See Figure 5 for plot.

(d) K=200 gives maximum accuracy of 59.3%. See Table 2

class	Precision	Recall
0	0.55	0.60
1	0.55	0.36
2	0.53	0.66
3	0.64	0.82
4	0.75	0.49

Table 2: Unweighted kNN implementation *d*

(e) Precision is 0 for all but class 0 (20%). Recall is 100% for class 0, since all instances are labeled with class 0. Matches expectations because of the tie-breaking rule.

(f) Same as above.

(g) Centroid baseline performs better than kNN. This implies that most data is concentrated around the centroids of the respective class, and is easily separable with linear decision boundaries. Any example where you presented multiple centroids for a single class will is probably be good example.

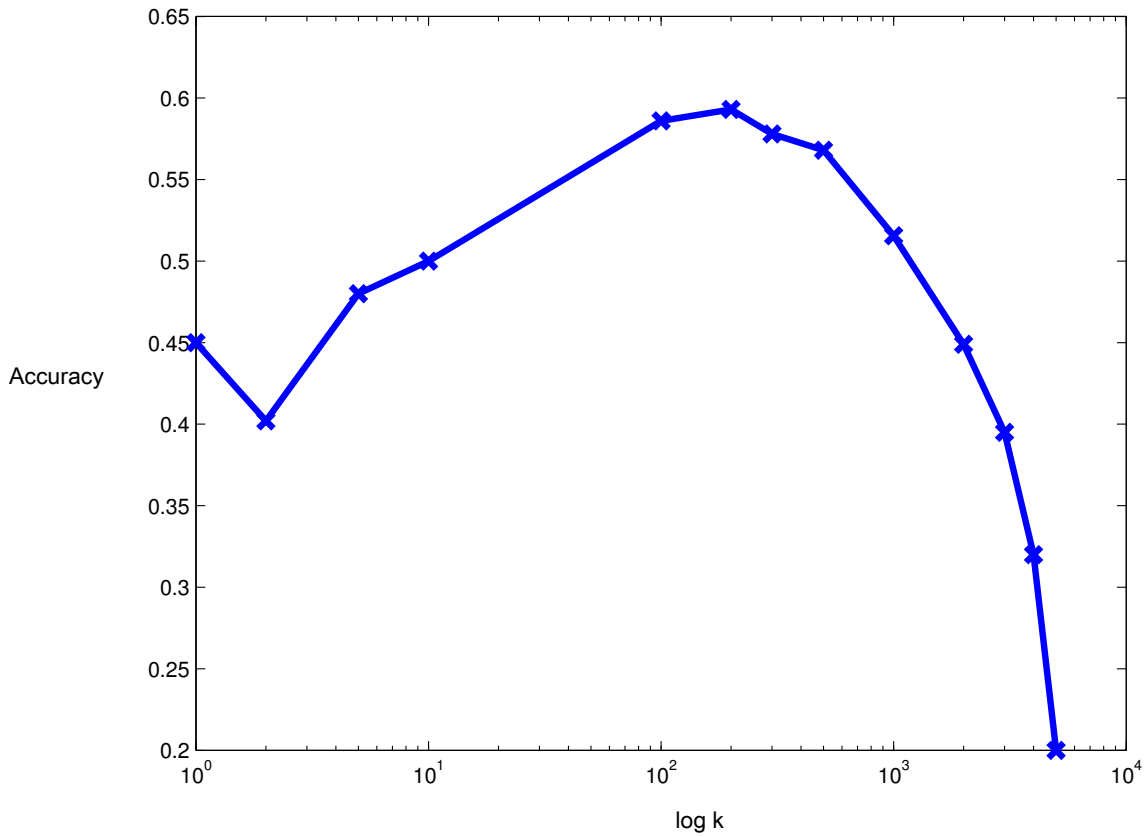


Figure 5: Unweighted KNN