Exam CS478

13th of April, 2004

Show all work on the exam. Colored paper is provided as scrap paper.

Exam is closed book and closed notes. Web and email may not be used.

Calculator with programming capabilities are not permitted.

NAME: ----------

I certify that the following work is my own

SIGNATURE: ----------

1 Bayes Rule

1. State Bayes Rule as a formula! (Note: I do NOT mean Bayes Theorem, i.e. $P(A|B) = P(B|A)P(A)/P(B)$)

5 pts.

2. State Bayes Rule as an english sentence!

5 pts.

3. Assume that you know for a given feature vector $\vec{x}$ and a two-class classification problem that $P(Y = 1|X = \vec{x}) = 0.2$ and $P(Y = -1|X = \vec{x}) = 0.8$. What is the probability of making an error on this example if you classify it as $y = 1$? What is the probability of error if you classify it as $y = -1$?

5 pts.

2 Version Spaces

1. A version space describes a set of hypotheses. How is this set defined?

   5 pts.

2. Consider the hypothesis space that consists of circles in two dimensions. A circle classifies all examples inside it (or on the border) as positive, while examples outside are negative. Draw four hypotheses so that h_1 is (strictly) more specific than h_2. h_3 is (strictly) more general than h_1, but h_3 is NOT more specific than h_2 and h_3 is NOT more general than h_2. h_4 is NOT more specific than h_2, but (strictly) more specific than h_3. Make sure you label the circles you draw!

   5 pts.

3 Decision Tree Learning

For the following questions, assume a top-down decision tree learning algorithm using information gain as we discussed it in class.

1. Correct or false: The ordering of the examples influences the decision tree that the algorithm returns. Justify your answer!

   3 pts.

2. Correct or false: If the examples have N attributes (all of which are binary), then the maximum number of nodes (with attribute tests) in a learned tree is $2^N - 1$. Justify your answer!

   3 pts.

3. Correct or false: Assume that there is at least one decision tree h with $Err_S(h) = 0$. The decision tree learning algorithm returns a tree that is consistent with the data and that has the minimum number of nodes.

   3 pts.

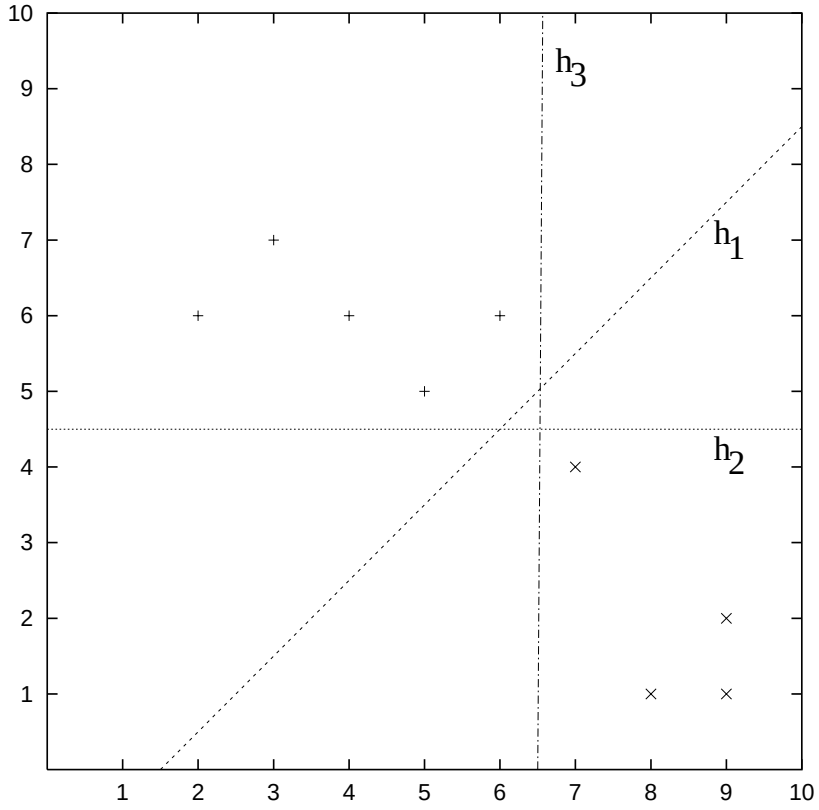
4 Perceptrons

1. Imagine the following situation: You train a perceptron (with $b = 0$) on a dataset and you notice that it made e updates. You know that all feature vectors $\vec{x}_i$ in the training set have Euclidean length one. What can you conclude about the (geometric) margin δ?

5 pts.

5 Optimal (Hard-Margin) Hyperplanes

The following graph has 9 points and three hyperplanes plotted into a coordinate system. "+" stands for a positive example, "x" for a negative example.



1. Which of the three hyperplanes is the maximum-margin hyperplane? Justify your answer by giving the definition of "maximum-margin hyperplane"!

3 pts.

2. Which of the points are support vectors? Mark them with "1" in the plot!

3 pts.

3. Which of the points are guaranteed to not be misclassified in leave-one-out testing? Mark them with “2” in the plot! Justify your answer with a one sentence explanation!

3 pts.

4. Indicate in the plot which distances are - by definition of maximum-margin hyperplanes - exactly equal to the size of the margin!

3 pts.

5. Given is a training sample $S = ((\vec{x}_1, y_1), ..., (\vec{x}_n, y_n)$ with $\vec{x}_i \in \Re^N$ and $y_i \in \{-1, +1\}$. Transform each vector $\vec{x}_i = (x_1, ..., x_N)$ to a vector $\vec{x}'_i = (x_1, ..., x_N, 0, ..., 0, 1, 0, ...0)$ with the 1 in the $(N+i)$-th position. Prove that this construction always leads to a linearly separable dataset (ie. $S = ((\vec{x}'_1, y_1), ..., (\vec{x}'_n, y_n)$ is linearly separable)!

10 pts.

6 Soft-Margin Support Vector Machines

1. Discuss two reasons for using a soft-margin SVM instead of a hard-margin SVM!

   6 pts.

2. How do training error and margin of a soft-margin SVM (generally) change for different values of C?

   6 pts.

   As a reminder, here is the primal soft-margin SVM optimization problem:

$$
\begin{aligned}
min_{\vec{w},b\vec{\xi}} \quad & \frac{1}{2}\vec{w}\cdot\vec{w} + C\sum_{i=1}^{n}\xi_i && (1)\\
s.t. \quad & y_1(\vec{w}\cdot\vec{x}_1 + b) \geq 1 - \xi_1 \text{ and } \xi_1 \geq 0 && (2)\\
& \ldots && (3)\\
& y_n(\vec{w}\cdot\vec{x}_n + b) \geq 1 - \xi_n \text{ and } \xi_n \geq 0 && (4)
\end{aligned}
$$

7 Kernels

Let $K(\vec{x}_i, \vec{x}_j)$ be a kernel so that for all $\vec{x}_i$ and $\vec{x}_j$: $K(\vec{x}_i, \vec{x}_j) = \phi(\vec{x}_i) \cdot \phi(\vec{x}_j)$. $\phi(\vec{x})$ is called the image of $\vec{x}$.

1. Name two reasons for why it might be beneficial to use a kernel!

   6 pts.

2. Given two points $\vec{x}_i$ and $\vec{x}_j$ in input space, show how you can compute the Euclidian distance between their images in feature space $\phi(\vec{x}_i)$ and $\phi(\vec{x}_j)$ without computing ϕ explicitly.

   6 pts.

8 Generative Classifiers

Explain the difference between generative and discriminative learning!

3 pts.

Name one generative and one discriminative learning algorithm!

2 pts.

9 Learning Theory

You have a learning algorithm that learns 1-out-of-m classifiers for an instance space with N binary features. A 1-out-of-m classifier is defined by a subset of $m \leq N$ features. If at least one of these m features is 1 for a given example, then the classifier outputs $y = 1$. Otherwise, it classifies it as $y = -1$.

1. For $m = 3$, your algorithm found a set of m features so that the 1-out-of-m classifier has zero training error on a sample of $n = 500$ examples. To be able to bound the prediction error via the general bound $P(Err_P(h_{\mathcal{L}}) \geq \epsilon) \leq |H|e^{-\epsilon n}$, you need to compute $|H|$. What is $|H|$ for 1-out-of-m classifiers in an instance space of N features? Explain your result!

   5 pts.

2. Show that you can construct an equivalent linear classfier for every 1-out-of-m classifier.

   5 pts.