IFT3395 / IFT6390
Fondements de l'Apprentissage Machine

Midterm Exam

Allowed documentation: 2 two-sided sheet of paper (letter format 8" 1/2 x 11") with your own course summary.

First Name:	**Professor:** Aaron Courville
Last Name:	**Assistants:** Philippe Lacaille
Code Permanent:	Tegan Maharaj
IFT3395 or IFT6390:	**Date:** October 18th, 2017
Study program (and lab if any):	

Exam is 100pts. Write directly in the spaces left blank. Answers should be concise yet precise. **Good luck!**

Notation

For all questions, we suppose we are working with a starting dataset containing n examples denoted $D_n = \{z^{(1)}, \ldots, z^{(n)}\}$. In the supervised case, $z^{(k)} = (x^{(k)}, y^{(k)})$ where $x^{(k)} \in \mathbb{R}^d$ is the input and $y^{(k)}$ is the corresponding target. Additional notation may be specified in the exam questions. You must respect the notations specified within this exam (adapt known equations as needed).

1 Work situation (15 pts)

You are working as a machine learning consultant to a company that wants to use machine learning to automate certain processes. They want a classifier capable of recognizing faces of authorized persons. The company is asking you for some advice about the dataset. They tell you the data consists of only images of the faces of authorized persons.

(a) Briefly describe a strategy to use this sort of data to accomplish the desired goal.

(b) Specify the form of learning problem you have defined (e.g. supervised learning, unsupervised learning, reinforcement learning, etc.)

(c) Now consider using a histogram-based method such as we've covered in the course. How you would apply this algorithm to train the classifier?

(d) For the histogram considered above, would you recommend using – as input to the algorithm – the images themselves (at a resolution of 256 pixels x 256 pixels) or would you recommend using a set of features taken from the images? Explain your answer.

2 Over-fitting, under-fitting, capacity and model selection (10 pts)

Reminder: the "capacity" of a machine learning algorithm corresponds, informally, to the "size" or "richness" or "complexity" of the considered set of functions among which it searches for the best prediction function.

Answer T for True or F for False (or leave a blank) to the left of each of the following statements: +1 for a correct answer, -1 for a wrong answer, 0 for an abstention (It's thus better not to answer a question you are unsure about. Minimum for the exercise is 0/10, maximum is 10/10).

(a) The more data examples we have, the more confidence we will have in using a high-capacity model without fear of overfitting.

(b) For the k-nearest neighbour classifier, the more nearest neighbours we use, the greater the capacity of the model.

(c) For the linear SVM, the number of support vectors is a hyperparameter which controls the capacity of the model.

(d) In practice, we choose our hyperparameters and report our generalization performance based on the performance on the test set.

(e) We diagnose overfitting by observing an increase in validation error while the training error continues to decrease over time.

(f) A model with high bias and low variance is a model that is more likely to overfit than underfit the data.

(g) The nearest centroid classifier (classifieur de plus proche moyenne) has more capacity than the perceptron classifier.

(h) The 1-nearest neighbour classifier has more capacity than the perceptron classifier.

(i) Cross-validation is a process of having one learning algorithm validate a second learning algorithm.

(j) In general, the more hyperparameter values we evaluate with validation set performance, the more biased this measure will be as an estimate of the true generalization performance.

3 Bias and Variance (12 pts)

For each case below, provide (i) a description of the phenomenon in terms of expected training set and test set performance; (ii) an example of a scenario where a learning algorithm could exhibit this pattern of behaviour. Be brief but specific. Include in your answer the desirability of this scenario and what can be done to ameliorate it if necessary.

(a) low bias, low variance

(b) high bias, low variance

(c) low bias, high variance

(d) high bias, high variance

4 Parzen Windows Regression (9 pts)

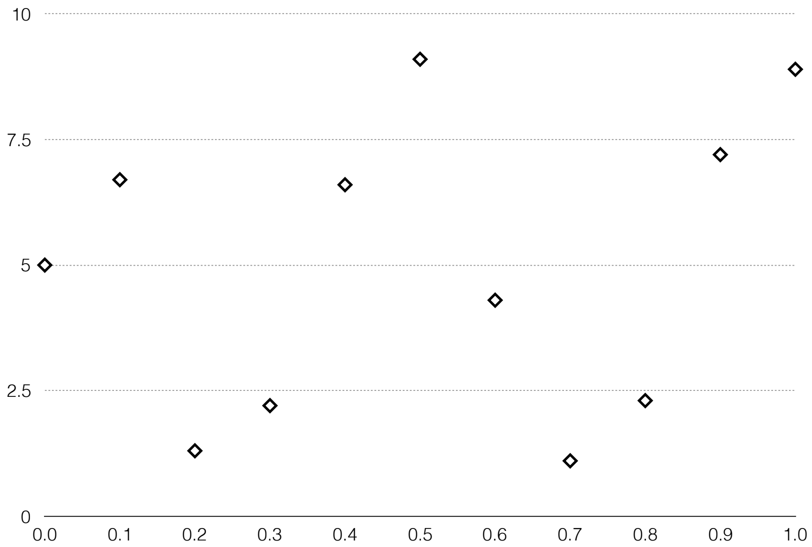
Consider a training dataset containing n examples denoted $D_n = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$ where $x^{(k)} \in \mathbb{R}^d$ is the input and $y^{(k)} \in \mathbb{R}$ is the corresponding regression target.

(a) Write the mathematical expression for the Parzen windows regression function with a Gaussian kernel of the form:

$$K(x, x^{(i)}) = \quad = \frac{1}{(2\pi)^{\frac{d}{2}} h^{d/2}} \exp\left(-\frac{1}{2h} \sum_{j=1}^{d} (x_j - x_j^{(i)})^2 \right)$$

(b) In your own words explain the difference between the Parzen windows regression algorithm and the k-nearest neighbours regression algorithm.

(c) On the following plot, draw two (approximate) regression predictions of your Parzen Windows regression function:



(i) With a solid line, for the kernel width $h = 0.05$.

(ii) With a dashed line, for the kernel width $h = 0.5$.

Clearly label the lines as either $h = 0.05$ or $h = 0.5$.

5 Parzen Windows and Bayes Classifiers (8 pts)

In this question, we consider a training dataset containing n examples denoted $D_n = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$ where $x^{(k)} \in \mathbb{R}^d$ is the input and $y^{(k)} \in \{1, \dots, c, \dots, M\}$ is the corresponding multi-class target.

You are going to explore a Bayes Classifier using a Parzen Windows density estimator with a Gaussian kernel of the form:

$$K(x, x^{(i)}) = \quad = \frac{1}{(2\pi)^{\frac{d}{2}} h^{d/2}} \exp\left(-\frac{1}{2h} \sum_{j=1}^{d} (x_j - x_j^{(i)})^2 \right)$$

We will consider the class prior $P(Y = c) = \gamma_c$, with $\sum_c \gamma_c = 1$ as known.

(a) Write the detailed expression for $P(x \mid Y = c)$.

(b) Write the detailed expression for $P(Y = c \mid x)$.

(c) Is this a naive Bayes classifier? Explain your answer.

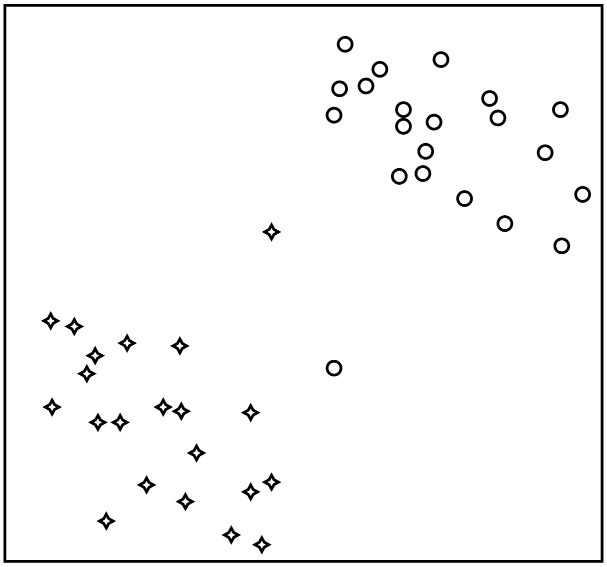
6 Linear SVM (12pts)

Consider a linear SVM classifier with the slack variables set to zero (i.e. hard-margin).

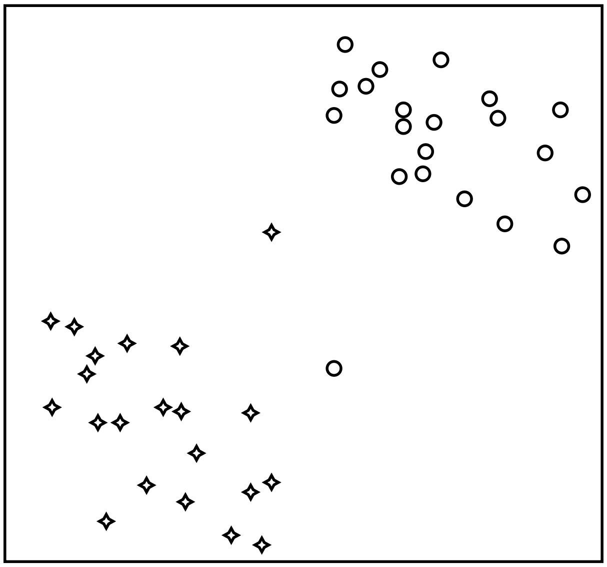
(a) What is the criterion that is optimized in determining the placement of the SVM decision boundary? Be specific.

(b) Argue that the distance between the decision boundary and the support vectors of the two classes is always the same (Hint: use a proof by contradiction).

(c) On the plot below, draw the SVM decision boundary that would result from training on the 2-dimensional dataset. Separate classes are indicated by different symbols. Circle the resulting support vectors.



(d) If we now allow for non-zero slack variables (i.e. soft-margin) how might the decision boundary change? Draw a possible solution below. Circle the resulting support vectors.

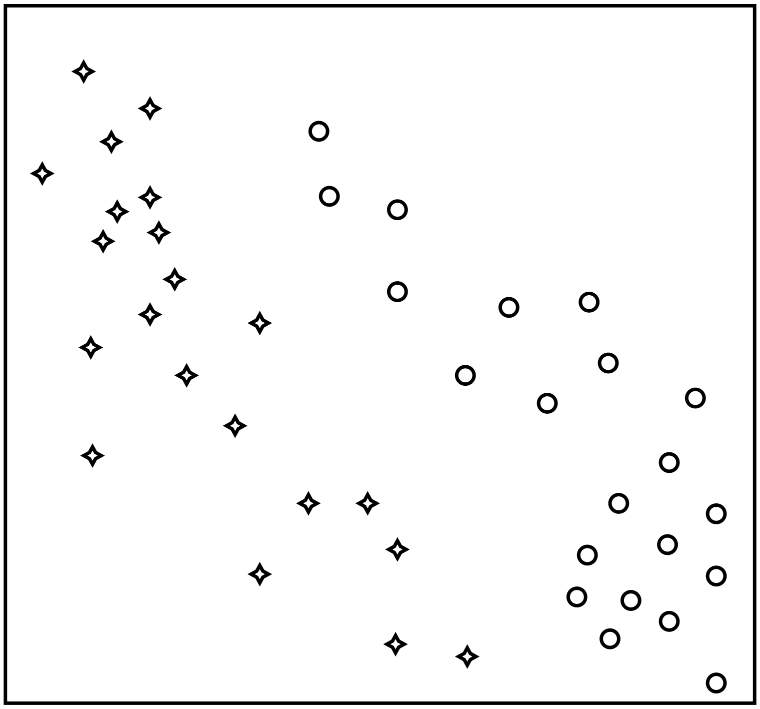


7 Gradient Descent and Logistic Regression (18pts)

(a) Consider a training dataset containing n examples denoted $D_n = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$ where $x^{(k)} \in \mathbb{R}^d$ is the input and $y^{(k)} \in \{0, 1\}$ is the corresponding binary classification target. For the case of logistic regression, where $g(x^{(i)}) = P(y^{(i)} = 1 \mid x^{(i)})$ is trained via minibatch gradient descent by using using the cross-entropy loss function, $L(g(x), y) = -(y \ln g(x) + (1 - y) \ln(1 - g(x)))$. For a mini-batch of size n', derive the parameter update rule for mini-batch gradient descent for this model.

(b) Write the training algorithm in pseudo-code for the algorithm derived in the previous question (i.e. part (a)).

8 Logistic Reg. vs Nearest Centroid Classifiers (6 pts)



(a) For the plot above, draw an "x" at the centroid of each of the two classes represented.

(b) Using a dashed line, draw the decision boundary for the Nearest Centroid classifier.

(c) Using a solid line, draw the decision boundary for the Logistic Regression classifier.

9 Maximum likelihood (10 pts)

(a) Consider a training dataset containing n examples denoted $D_n = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$ where $x^{(k)} \in \mathbb{R}^d$ is the input and $y^{(k)} \in \{0, 1\}$ is the corresponding binary classification target. We take the output of our training algorithm to be $g(x^{(i)}) = P(y^{(i)} = 1 \mid x^{(i)}))$. Show with the cross-entropy loss function, $L(g(x), y) = -(y \ln g(x) + (1 - y) \ln(1 - g(x)))$, the empirical risk can be naturally derived from maximum likelihood principle on a Bernoulli distribution, shown below with parameter p.

$$\text{Bernoulli}(y; p) = p^y (1 - p)^{(1-y)}$$

(b) Now consider a training dataset containing n unlabeled examples denoted $D_n = \{x^{(1)}, \ldots, x^{(n)}, \}$ where $x^{(k)} \in \mathbb{R}$. Derive the maximum likelihood parameter estimator for the β parameter for the Gamma probability density function (PDF), given below. Assume the parameter α is known.

$$\text{Gamma}(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(\beta x)$$