Introduction to Supervised Learning, COMPGI20, 2016

3 Hours

No electronic devices are acceptable, including calculators and smartphones.

Answer all questions

For all True/False questions: always support your reply with a short answer, using at most three sentences, or approximately 10-50 words in total (whatever suits you best). You may find it easier to reply using equations - in that case there is no sentence, or word count. If you do not provide a justification, your answer will not be taken into consideration, whether true or false.

Marks for each part of each question are indicated in square brackets

Calculators are NOT permitted

Probability

1. Source: Kevin Murphy's book on machine learning, Exercise 2.2.

   Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1

   - The prosecutor claims: There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus there is a 99% chance that he guilty. This is known as the prosecutors fallacy. What is wrong with this argument?
   - The defender claims: The crime occurred in a city of 800,000 people. The blood type would be found in approximately 8000 people. The evidence has provided a probability of just 1 in 8000 that the defendant is guilty, and thus has no relevance. This is known as the defenders fallacy. What is wrong with this argument?

Linear Regression, Maximum Likelihood Estimation

We have a dataset with R records in which the i^{th} record has one real-valued input attribute x_i and one real-valued output attribute y_i.

(a) *(6 points)* First, we use a linear regression method to model this data. To test our linear regressor, we choose at random some data records to be a training set, and choose at random some of the remaining records to be a test set.
Now let us increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training and mean testing errors? (No explanation required)

- Mean Training Error: A. Increase; B. Decrease

- Mean Testing Error: A. Increase; B. Decrease

(b) *(6 points)* Now we change to use the following model to fit the data. The model has one unknown parameter w to be learned from data.

$$y_i \sim N(log(wx_i), 1)$$

Note that the variance is known and equal to one. (no explanation required) Suppose you decide to do a maximum likelihood estimation of w. You do the math and figure out that you need w to satisfy one of the following equations. Which one?

A. $\sum_i x_i \log(wx_i) = \sum_i x_i y_i \log(wx_i)$

B. $\sum_i x_i y_i = \sum_i x_i y_i \log(wx_i)$

C. $\sum_i x_i y_i = \sum_i x_i \log(wx_i)$

D. $\sum_i y_i = \sum_i \log(wx_i)$

Generalization, Regularization, Loss Functions

a. Are the following statements true or false? Justify your response.

- A classifier trained on less training data is less likely to overfit.
- It is always better to use models with fewer parameters.
- Increasing the value of the regularizer always increases the generalization performance.
- The Logistic loss is better than the L2 loss in classification tasks

Support Vector Machines

2. Consider a supervised learning problem in which the training examples are points in 2-dimensional space. The positive examples are at positions $(1,1)$ and $(-1,-1)$. The negative examples are at positions $(1,-1)$ and $(-1,1)$.

   a. Are the positive examples linearly separable from the negative examples in the original space? Show this visually using a 2D layout of your data.

   b. Consider the feature transformation $\phi(x) = [1, x_1, x_2, x_1 x_2]$, where $x = (x_1, x_2)$, i.e. x_1 and x_2 are, respectively, the first and second coordinates of a generic example x. The prediction function is $y(x) = w^T \phi(x)$ in this feature space. Give the coefficients, w, of a maximum-margin decision surface separating the positive examples from the negative examples. Do this by inspection, without any computation.

   [2 marks]

   c. What kernel $K(x, x')$ does this feature transformation ϕ correspond to?

   [2 marks]

Boosting

Consider a text classification task, such that the document X can be expressed as a binary feature vector of the words. More formally $X = [X_1, X_2, X_3, ..., X_m]$, where $X_j = 1$ if word j is present in document X, and zero otherwise. Consider using the AdaBoost algorithm with a simple weak learner, namely

$$h(X; \theta) = yX_j$$
$$\theta = \{j, y\} \; j \text{ is the word selector ; } y \text{ is the associated class}$$
$$y \in \{-1, 1\}$$

More intuitively, each weak learner is a word associated with a class label. For example if we had a word `football`, and classes {`sports`,`non-sports`}, then we will have two weak learners from this word, namely

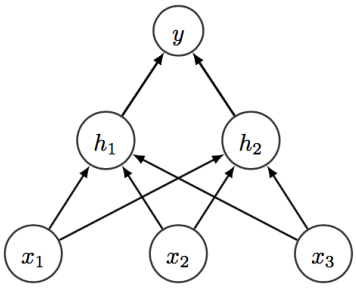
- *Predict* **sports** *if document has word* **football**
- *Predict* **non-sports** *if document has word* **football**.

d.
- How many weak learners are there ?
- This boosting algorithm can be used for feature selection. We run the algorithm and select the features in the order in which they were identified by the algorithm. Can this boosting algorithm select the same weak classifier more than once? Explain.

e.
- true or false? AdaBoost will eventually give zero training error regardless of the type of weak classifier it uses, provided enough iterations are performed.

Neural Networks

The following graph shows the structure of a simple neural network with a single hidden layer. The input layer consists of three dimensions $x = (x_1, x_2, x_3)$. The hidden layer includes two units $h = (h_1, h_2)$. The output layer includes one unit y. We ignore bias terms for simplicity.



We use linear rectified units $\sigma(z) = \max(0, z)$ as activation function for the hidden and the output layer. Moreover, denote by $l(y,t) = \frac{1}{2}(y-t)^2$ the loss function. Here t is the target value for the output unit y. Denote by W and V weight matrices connecting input and hidden layer, and hidden layer and output respectively. They are initialized as follows:

$$W = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} \text{ and } V = \begin{bmatrix} 0 & 1 \end{bmatrix} \text{ and } x = [1,2,1] \text{ and } t = 1.$$

Also assume that we have at least one sample (x,t) given by the values above.

- Write out symbolically (no need to plug in the specific values of W and V yet) the mapping $x^T y$ using σ, W, V.
- Assume that the current input is $x = (1,2,1)$. The target value is $t = 1$. Compute the numerical output value y, clearly showing all intermediate steps. You can reuse the results of the previous question.
- Compute the gradient of the loss function with respect to the weights. In particular, compute the following terms symbolically:
  - The gradient relative to V , i.e. $\frac{\partial l}{\partial V}$
  - The gradient relative to W, i.e. $\frac{\partial l}{\partial W}$
  - Compute the values numerically for the choices of W, V, x, y given above.

New exercises

Linear Regression

3. a. What does it mean for a regression problem to be ”under-determined” and what does it mean to be ”over-determined”?

[2 marks]

b. Consider that you are provided with an under-determined problem. Describe how you can avoid over-fitting and choose any additional parameters that may need to be set.

[2 marks]

SVMs

4. Some general questions relating to the properties of SVMs.

a. Consider a point that is correctly classified and distant from the decision boundary. Why would the SVMs decision boundary be unaffected by this point, but the one learned by logistic regression be affected?

[4 marks]

b. Why does the kernel trick allow us to solve SVMs with high dimensional feature spaces, without significantly increasing the running time?

Adaboost

5. The Adaboost training algorithm is provided below for your reference:

   Initially, set $D_1(i) = \frac{1}{N}, \quad \forall i$

   For $t = 1 \ldots T$:

   - Find weak classifier $h_t : \mathcal{X} \to \{-1, 1\}$ with smallest weighted training error

   $$\varepsilon_t = \frac{\sum_{i=1}^{N} D_t^i [y^i \neq h_t(x^i)]}{\sum_i D_t^i} \tag{1}$$

   - Set $a_t = \frac{1}{2} \log \frac{(1-\varepsilon_t)}{\varepsilon_t}$
   - Update distribution

   $$\begin{aligned} D_{t+1}^i &= \frac{D_t^i}{Z_t} \times \begin{cases} \exp(-\alpha_t), & \text{if} y^i = h_t(x^i) \\ \exp(\alpha_t), & \text{if} y^i \neq h_t(x^i) \end{cases} \\ &= \frac{D_t^i}{Z_t} \exp(-\alpha_t y^i h_t(x^i)), \end{aligned}$$

   where to guarantee that D_{t+1} remains a proper density we normalize by:

   $$Z_t = \sum_i \frac{D_t^i}{\exp}(-\alpha_t y^i h_t(x^i))$$

   - Combine classifier results (weighted vote)

   $$f(x) = \sum_t a_t h_t(x)$$

   - Output final classifier:

   $$H(x) = \text{sign}\,(f(x)) \tag{2}$$

You are asked to answer the following TRUE/FALSE questions regarding this algortihm and justify your response.

a. The weighted training error ε_t of the t-th weak classifier on training data with weights D^t tends to increase as a function of t.

[5 marks]

b. The votes a_t assigned to the classifiers assembled by AdaBoost are always non-negative.

[5 marks]

END OF PAPER