Machine Learning, IUB Fall 2004, Midterm Exam

-- With selected solutions --

October 8, 2004, 14:00

Problem 1. Assume you want to predict the oil price O(t+1) for day t+1 (= tomorrow) from the oil prices O(t), O(t–1), ... recorded at days t (= today), t–1, t – 2, ... In addition, you have records **x**(t), **x**(t–1), ... of 80 other oil-market relevant economy indicators collected in 80-dim vectors **x.** You wish to train a predictor π that gets as input whatever you decide is useful from the known data up to day t (today), and returns an estimate of tomorrow's oil price.

1. (10 points) Frame this situation in terms of an event space Ω and suitable random variables and their measure spaces. There are many plausible candidates for Ω, the choice is yours. Describe the elements of your Ω in words (you know they can't be specified formally), and think out two concrete examples for relevant random variables, together with their measure spaces, one numerical and one consisting of a finite number of qualitative classes.
2. (5 points) Any experienced machine learning expert would pre-process the raw data **x**(t) before entering them into a learning procedure. What is the general purpose of such preprocessing?
3. (10 points) You know about the bias-variance dilemma in general. Give reasons why it is likely to be particularly relevant for this prediction problem.

Problem 2. Consider a 2-class classification problem with training samples (x_i, y_i), where the x_i are just points on the real line, and the $y_i \in$ {A,B} are class labels. An input x belongs to class A if -1 < x < 1, else it belongs to class B. The patterns x are uniformly distributed over the interval Ω = [-3,+3].

1. (2 points) What is the domain and the image of the random variable Y that corresponds to this situation?
2. (2 points) What is the numerical value of $P(Y = \text{A})$?
3. (10 points) Consider the linear discriminant $y(x) = w_1 x + w_0$, which decides for class A if $y(x) \geq 0$. Which values for w_1 and w_0 yield the lowest rate r of misclassifications $r = P(y(x) \geq 0 \mid Y(x) = B) + P(y(x) < 0 \mid Y(x) = A)$?
4. (10 points) Find a function (feature) $F(x)$ and values for w_1 and w_0 such that the linear discriminant $y(x) = w_1 F(x) + w_0$ has zero misclassification rate.

Solution.

1. domain(Y) = [-3,+3], image(Y) = {A,B}.
2. $P(Y = \text{A}) = 1/3$.

3. The decision function $y(x) = w_1 x + w_0$ is a straight line with slope w_1 that crosses the x-axis at $d = - w_0/w_1$ (if w_1 is not equal to 0). By a case distinction or drawing, it is easy to see that the rate of misclassification is always at least 1/3, except for the following cases:
    a. $- w_0/w_1 \leq -3$ and $w_1 < 0$, (in this case all x from class A are misclassified)
    b. $- w_0/w_1 \geq 3$ and $w_1 > 0$, (again, exactly the x from class A are misclassified)
    c. $- w_0/w_1 = -1$ and $w_1 > 0$, (only the $x > 1$ are misclassified as A's)
    d. $- w_0/w_1 = 1$ and $w_1 < 0$, (only the $x < -1$ are misclassified as A's)
    e. $w_1 = 0$ and $w_0 < 0$ ("degenerate" case, no decision boundary, only the x from class A are misclassified).

Any of these cases would count as a solution.

4. One feature that works is $F(x) = - x^2$. Then if $d = - w_0/w_1 = -1$, values of $F(x)$ larger than –1 lead to an "A" decision. But $F(x) = - x^2 > -1$ iff $|x| < 1$, that is, iff x is in class A. So we always get correct classifications with $F(x) = - x^2$ and again any nonzero $w_0 = w_1$.

Problem 3. A *dynamic pattern* is a sequence of 2-d pixel images – you may also call it a movie clip. A dynamic pattern is *stochastic* if the n+1-th image cannot be determined by a deterministic function from the n-th image. A dynamic pattern is *time-stationary* if its properties don't change over time, that is, if you are shown some subsequence you couldn't tell whether it comes from the beginning or the end of the entire sequence (e.g., a movie showing a gas burner flame run with constant gas supply would be time-stationary, while a movie showing a candle burning down would not be time-stationary). A dynamic pattern is *space-stationary* if the local visual dynamics are the same at all positions within the pixel field (e.g., a movie showing waves on a patch of water would be space-stationary). Your task: design a "trainable retina", that is, a system that is capable to learn stochastic time- and space-stationary dynamic patterns from a presentation of a training dynamic pattern. The trainable retina should have an input retina (same pixel size as training patterns) and an output retina (same size again). After training, the system should be able to generate on its output retina dynamic patterns of the same kind as seen during learning. Hints:

1. I don't expect a complete treatment but want to see how you can develop relevant ideas. Don't spend more than 20 minutes / 1.5 pages on this.
2. What I want to see minimally is *some* idea for a design of a trainable retina that *might* work (there are many possibilities). The main idea should be described as precisely and succinctly as possible. (20 points)
3. In addition, comment on what tricky issues you expect to encounter if you would actually implement and optimize such a trainable retina. What properties of training patterns would facilitate learning or render it difficult? How could the quality of such a device be measured? what about "temporal memory", that is, can it occur that pattern n is influenced not only by pattern n-1 but also by pattern n-2, n-3, ... ? What about

boundary effects? Long-term stability of autonomously generated dynamic patterns? Bias-variance dilemma? and many more... (20 points)

Problem 4. (25 points) Show strong consistency of θ_i^{PME}, the Bayesian posterior mean estimator from Section 3.3. of the lecture notes. Hint: follow the lines of the similar proof of strong consistency of θ_i^{ML} in Section 3.4 of the lecture notes.

Solution:

For a sample of size N, ${\theta_{i,N}}^{\mathrm{PME}}(\omega) = \dfrac{a_i + \sum_{j=1}^{N} X_j(\omega)}{A+N}$, where $X_j(\omega) = 1$ if the j-th protein sequence in our sample has the i-th amino acid symbol in the location of interest, else $X_j(\omega) = 0$. The X_j are integrabel, independent, and identically distributed, so the strong law applies. The expectation of X_j is θ_i for all j. So we can conclude:

$$
\begin{aligned}
P(\lim_{N\to\infty} \theta_{i,N}^{PME}(\omega) = \theta_i) &= P(\lim_{N\to\infty} \frac{a_i + \sum_{j=1}^{N} X_j(\omega)}{A+N} = \theta_i) \\
&= P(\lim_{N\to\infty} \frac{a_i}{A+N} + \frac{\sum_{j=1}^{N} X_j(\omega)}{A+N} = \theta_i) \\
&= P(\lim_{N\to\infty} \frac{\sum_{j=1}^{N} X_j(\omega)}{A+N} = \theta_i) \\
&= P(\lim_{N\to\infty} \frac{\sum_{j=1}^{N} X_j(\omega)}{N} = \theta_i) \\
&= \ldots \\
&= 1
\end{aligned}
$$

where the "..." stand for the identical lines from the lecture note for the case of ${\theta_i}^{\mathrm{ML}}$.