Advanced Machine Learning, IUB Fall 2004, Final Exam -- Solutions

Problem 1. (20 points) After one semester of "Advanced Machine Learning" you know a lot more about machine learning than you did before – in other words, you have learnt something. Explain in words **(a)** one aspect of this 1-semester personal learning experience that can be understood as supervised learning, **(b)** another aspect that demonstrates mechanisms of unsupervised learning, **(c)** yet another one that cannot readily be understood as either. Target size for each explanation: 100 words.

(no particular "correct" solution)

Problem 2. (20 points) Earlier today I had a strange experience that apparently contradicts common machine learning wisdom. On a 1000-step long symbol sequence which was generated by a very complex stochastic sequence generation mechanism, I trained (i) a 5-state HMM and (ii) a 3-dimensional OOM. The correct log probability of the training sequence (known only to me) was – 313.5; the log probability assigned to the training data by the trained models were – 314.2 for the HMM and – 324.0 for the OOM. This led me to believe that the HMM would be the better model, because, so I reasoned, in the sense of maximum likelihood estimation it had a better likelihood while not overfitting. However, on independent test data the OOM model turned out to be superior by a large margin. Is there a fault in the belief that the HMM did not overfit, or can you offer another explanation? (Target size of answer: 120 – 180 words.)

Solution. The HMM *did* overfit. Generally speaking, if a model assigns to the training data a probability that is lower than the correct one, this is no safeguard against overfitting. Intuitively speaking, the HMM was the combined result of two phenomena: (i) mere data fitting without capturing the underlying rule of the data, (ii) inability to capture (parts of) the underlying rule. Aspect (i) leads to a high model likelihood, (ii) to a low one; the two effects can in principle mix in any ratio. Thus, the HMM "modeled successfully" some random noise contribution to the data while structurally being unable to capture some important aspects of the original dynamics. The OOM was better in the latter respect but not high-dimensional enough to fall prey to overfitting as badly as the HMM.

Problem 3. (25 points) Consider the linear system with the update equation $x(n) = 0.1\ u(n) + u(n-1)$, where $u(n)$ is a white noise zero mean input signal with variance $\sigma_u^2 = 1.0$. This system is used to generate a training sequence $x(1), x(2), \ldots$. The LMS algorithm is used to train a predictor filter of the form $\hat{x}(n+1) = w_0\ x(n) + w_1\ x(n-1)$.

- **(a)** (15 points) Compute the correlation matrix **R** and the correlation vector **p** that would be needed to obtain the optimal filter weights w_0, w_1 from the Wiener-Hopf equation.
- **(b)** (10 points) Give a rough numerical estimate of the convergence rate that LMS can achieve on this task when the learning rate is optimally selected.

Solution. (a) The variance of $x(n)$ is $\mathrm{var}(x(n)) = \mathrm{var}(0.1\ u(n)) + \mathrm{var}(u(n-1)) = 0.01\ \mathrm{var}(u(n)) + \mathrm{var}(u(n-1)) = 1.01$. Furthermore $E[x(n)\ x(n-1)] = E[(0.1\ u(n) + u(n-1))\ (0.1\ u(n-1) +$

$u(n-2))] = E[\, u(n-1)\; 0.1\, u(n-1)\,] = 0.1\, \sigma_u^{\,2} = 0.1$ and $E[x(n)\, x(n-2)] = E[(0.1\, u(n) + u(n-1))$ $(0.1\, u(n-2) + u(n-3))] = 0$. This gives $\mathbf{R} = E[(x(n)\, x(n-1))^{\mathsf{T}}\, (x(n)\, x(n-1))] =$ [1.01 0.1; 0.1 1.01] *(using Matlab notation)* and $\mathbf{p} = E[(x(n)\, x(n-1)); (x(n)\, x(n-2))] =$ [0.1; 0]. Incidentally, this gives optimal filter weights $\mathbf{w} = \mathbf{R}^{-1}\, \mathbf{p} =$ [0.1; -0.0099].

(b) LMS would converge quickly, because **R** is close to the identity matrix, that means, its eigenvalues will be both close to 1, that is, the two modes of convergence will not differ by much, that is, the optimal learn rate can be set close to 1/2 (cf. equation (6.30) from script), that is, the rate of convergence will be close to 2.

Problem 4. (20 points) An old gambler with bad eyesight suspects that the die he has recently bought is loaded. Unfortunately, when he sees the face of the die that has come up he knows that he might misread it (due to his weak eyes); even more unfortunately, he does not know the probabilities by which he misreads a certain true outcome for another, perceived outcome, that is, he has no clue whatsoever about the probabilities $q_{ij} = P$(perceived face is j | true face is i). All he can do is throw the die many times and record what he *thinks* he sees. So he throws the die 10,000 times and gets a sequence of observations $D = y(1), ..., y(10\,000)$...

- **(a)** (8 points) Describe the structure of a HMM for this process, by specifying a suitable set of hidden states and observable events. Assuming that the die displays the 6 faces with probabilities $p_1, ..., p_6$, what are the transition probabilities $p_{ii'}$ of your Markov transition matrix? What are the emission probabilties e_{ij}?
- **(b)** (8 points) The gambler dimly remembers from his young days as an IUB student that the EM algorithm can be used to estimate HMM parameters from an observation sequence. Could he indeed infer from D information about whether the die is loaded? If yes, how? if not, why would EM not work? Or could it work sometimes, but not in all cases?
- **(c)** (4 points) Describe the same process by an OOM. What is the minimal model (= process) dimension? Specify the observable operators in terms of the q_{ij} and p_i.

Solution. (a). There are at least two ways of how this process can be framed as an HMM. Version 1: we use six hidden states s_i, each of them corresponding to the hidden event "the true face is i". The emission probabilities e_{ij} are then the given q_{ij}. The transition probabilities $p_{ii'}$ are $p_{ii'} = p_{i'}$. Version 2: Use a single hidden state (which then has self-transition probability $p_{11} = 1$). For the emission probability of perceived face j put $e_{1j} = \sum_{i=1,...,6} p_i q_{ij}$.

(b) The gambler can *never* find out *anything* about whether the die is loaded or not. One way of formal reasoning to make this point is to consider the version 2 model from (a). It shows that the process is fully described by 6 parameters; these form a sufficient statistic for the process. If the gambler wanted to learn something about the p_i, he would have to be able to infer that from the e_{1j}. However, because no prior information is available about the q_{ij}, *any* distribution for the p_i would be compatible with any estimate of the e_{1j}.

(c) Because the process has no memory, all predictor functions $g_{\bar{a}}$ are identical; they span a one-dimensional vector space and thus the process dimension is 1. Thus, the observable operators $\tau_1, ..., \tau_6$ are one-dimensional linear maps, that is, multiplication scalars. They are $\tau_j = e_{1j}$.

Problem 5. (25 points) Let (Ω, $\mathcal{A}$, P) be a probability space. Prove $\forall A, B \in \mathcal{A}$: $A \subseteq B \to P(A) \leq P(B)$. Prove this from first principles, that is, start from the 5 axioms of Kolmogorov [remember they were 1. $\Omega \in \mathcal{A}$; 2. $\forall A \in \text{Pot}(\Omega)$: $A \in \mathcal{A} \Rightarrow A^{C} \in \mathcal{A}$; 3. for every sequence $(A_n)_{n = 1, 2, \ldots}$ in F, the set $\bigcup_{n=1}^{\infty} A_n$ is in F; 4. $P(\Omega) = 1$; 5. for every sequence $(A_n)_{n = 1, 2, \ldots}$ of pairwise disjoint events it holds that $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.]. Make sure that you don't use unproven arguments in your proof (e.g., don't use $P(A \cup B) = P(A) + P(B)$ [A, B disjoint] without proving it!)

Solution: We are pedantic here and really go through all necessary details of the proof. We start by two auxiliary claims.

Claim 1: $\varnothing \in \mathcal{A}$. Proof: follows from axioms 1. and 2.
Claim 2: $P(\varnothing) = 0$. Proof: assume $P(\varnothing) = a > 0$. Then consider the constant sequence $(A_n)_{n = 1, 2, \ldots} \equiv \varnothing$. It is a sequence of pairwise disjoint sets, and we can apply axiom 5 to conclude $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} P(\varnothing) = \sum_{n=1}^{\infty} a = \infty$, contradiction to the fact that the image of P is $[0,1]$.

Main argument: If $A \subseteq B$, then $B = A \cup A'$ with A disjoint from A'. Then $P(B) = P(A \cup A')$ $= P(\bigcup_{n=1}^{\infty} A_n)$, where we put $A_1 = A$, $A_2 = A'$, and $A_i = \varnothing$ for $i > 2$. By axiom 5 and claim 2, we get $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n) = P(A) + P(A')$. Because $P(A') \geq 0$, we get $P(B) = P(A') + P(A)$ $\geq P(A)$, and we are done.