

58147 Machine Learning (Spring 2005)

Exercise 8 (Wednesday 23 March)

1. In the following, you may use the result from Exercise 7, Problem 2, and its more general form that $\mathbb{E}[\mathbb{E}[X \mid \mathcal{F}_2] \mid \mathcal{F}_1] = \mathbb{E}[X \mid \mathcal{F}_1]$ for σ -algebras $\mathcal{F}_1 \subseteq \mathcal{F}_2$.

- (a) Consider n independent throws of a symmetrical six-sided die. Let Y_i be the result of the i th throw, and $Y = Y_1 + \dots + Y_n$. Define

$$X_i = \mathbb{E}[Y \mid Y_1, \dots, Y_i].$$

Write out X_i in a closed form (involving Y_1, \dots, Y_i). Show that

$$\mathbb{E}[X_{i+1} \mid Y_1, \dots, Y_i] = X_i.$$

- (b) More generally, fix a filter (\mathcal{F}_i) and a random variable Y , and define $X_i = \mathbb{E}[Y \mid \mathcal{F}_i]$. Show that (X_i) is a martingale with respect to (\mathcal{F}_i) . (Use result from Exercise 7, Problem 3.)
2. Consider a graph $G = (V, E)$, where $V = \{1, \dots, n\}$, and for $1 \leq i < j \leq n$ write $I_{ij} = 1$ if $(i, j) \in E$ and $I_{ij} = 0$ otherwise. Let $\chi(G)$ be the chromatic number of G , *i.e.*, the smallest number of colours needed to colour the vertices of G in such a way that no two adjacent vertices have the same colour.

Assume now that G is generated at random so that the I_{ij} are independent of each other, and $\Pr(I_{ij} = 1) = p$. (Here n and p are constants.) Let \mathcal{F}_k be the smallest σ -algebra such that I_{ij} is measurable for all $1 \leq i < j \leq k$. Define $X_i = \mathbb{E}[\chi(G) \mid \mathcal{F}_i]$.

Show that $|X_{i+1} - X_i| \leq 1$. Then use Azuma's inequality and previous problem to prove

$$\Pr(|\chi(G) - \mathbb{E}[\chi(G)]| > \lambda/\sqrt{t}) \leq 2 \exp(-\lambda^2/2).$$

(This is not directly related to learning, it's just a nice application of martingales.)

3. Let $A \in \mathbf{R}^{n \times n}$ be a matrix, and define $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T A \mathbf{z}$.
 - (a) Show that (\mathbf{R}^n, k) is an inner product space if and only if A is symmetrical and positive semi-definite.
 - (b) Assuming A is symmetrical and positive semi-definite, construct a "feature map" $\psi: \mathbf{R}^n \rightarrow \mathbf{R}^n$ such that $k(\mathbf{x}, \mathbf{z}) = \psi(\mathbf{x}) \cdot \psi(\mathbf{z})$ where \cdot is the usual dot product. (*Hint*: use the eigen-decomposition.)
4. Let $Z = \{1, \dots, n\}$ and $X = \mathcal{P}(Z)$ (the set of all subsets of Z). For $A, B \in X$, let $k_1(A, B) = 2^{|A \cap B|}$ be the number of common subsets A and B have. Assume additionally that P is a probability measure over Z , and let $k_2(A, B) = P(A \cap B)$.

Show that k_1 and k_2 are valid kernels. In both cases, do this by explicitly constructing a feature map and feature space. (*Hint*: consider 2^n -dimensional and n -dimensional feature spaces.) Which elements of the feature space are actually of the form $\psi(A)$ for some $A \in X$?

5. Let $X = \mathbf{R}^n$. Given a sample $S = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$, let $V = \{\sum_{i=1}^m \alpha_i \mathbf{x}_i \mid \alpha_i \in \mathbf{R}\}$ be the span of the example points. Let $C > 0$, and define

$$R(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^m (\mathbf{w} \cdot \mathbf{x}_i - y_i)^2 + \frac{C}{2} \|\mathbf{w}\|_2^2.$$

Show that the minimiser $\mathbf{w}_* = \arg \min R(\mathbf{w})$ is in V .

Hint: There is no need to solve the minimisation explicitly. Just take any $\mathbf{w} \in V$ and notice that adding a component orthogonal to all the \mathbf{x}_t will always increase $R(\mathbf{w})$.

Notice that the result generalises to very general classes of loss functions and *regularisers* (such as $\frac{C}{2} \|\mathbf{w}\|_2^2$ in the problem.)