

Exercise 8.1

a) Let Y_* be the expected value of a single throw. Now,

$$X_i = E[Y|Y_1, \dots, Y_i] = E[Y_1 + \dots + Y_n | Y_1, \dots, Y_i] \quad (112)$$

$$= Y_1 + \dots + Y_i + (n - 1)E[Y_*], \quad (113)$$

as the throws are independent. But also

$$\begin{aligned} E[X_{i+1}|Y_1, \dots, Y_i] &= E[Y_1 + \dots + Y_{i+1} + (n - (i + 1))E[Y_*]|Y_1, \dots, Y_i] \\ &= Y_1 + \dots + Y_i + E[Y_{i+1} + (n - i - 1)E[Y_*]|Y_1, \dots, Y_i] \\ &= Y_1 + \dots + Y_i + E[Y_{i+1} + (n - i - 1)E[Y_*]] \\ &= Y_1 + \dots + Y_i + E[(n - i)E[Y_*]] \\ &= Y_1 + \dots + Y_i + (n - 1)E[Y_*], \end{aligned}$$

due to independence assumptions and $E[Y_i] = E[Y_*], \forall i$.

b) For (X_t) to be a martingale sequence w.r.t. (\mathcal{F}_t) , we need to show $E[X_t | z_0, \dots, z_{t-1}] = X_{t-1}$ for all random sequences (Z_t) such that Z_t is \mathcal{F}_t measurable.

Due to exercise 7.3, it suffices to show that $E[X_{t+1} | \mathcal{F}_t] = X_t$. The wanted result follows from this. As we defined $X_t = E[Y | \mathcal{F}_t]$, insert this definition to the right side of the previous formula, yielding

$$E[E[Y | \mathcal{F}_{t+1}] | \mathcal{F}_t], \quad (114)$$

which is equal to

$$E[Y | \mathcal{F}_t] = X_t, \quad (115)$$

according to the more general version of 7.2 mentioned in the exercise sheet. We also need measurability of X_i w.r.t \mathcal{F}_i for this to work, but this follows from the definition of X_i . Hence, we are done.

□

Exercise 8.2

Here we are interested in proving that the expected number of colors required to color a graph G is concentrated around its expectation.

First we need to show

$$|X_{i+1} - X_i| = |E[\chi(G)|\mathcal{F}_{i+1}] - E[\chi(G)|\mathcal{F}_i]| \leq 1. \quad (116)$$

Do this by numbering the nodes of the graph arbitrarily. Examine node $i + 1$. Due to measurability of \mathcal{F}_{i+1} , this is equal to revealing information about the edges from the node $i + 1$ to the nodes with smaller index. Clearly adding all such $i + 1$ -ending possible edges to G , or removing all such edges from G can only increase or decrease the number of required graph colorings by at most 1 (the new node $i + 1$ either needs to have a new color than the rest when adding edges, or had an unique color, when removing edges).

Now, from (\mathcal{F}_i) being a filter sequence it follows that (X_i) is a martingale (previous ex.). We also now know that it fulfills the criteria $|X_{i+1} - X_i| < 1 = c_1$ as required by Azuma's inequality. Now apply that by writing

$$P(|X_t - X_0| \geq \alpha) \leq 2 \exp\left(\frac{-\alpha^2}{2t}\right) \quad (117)$$

$$P(|E[\chi(G)|\mathcal{F}_t] - E[\chi(G)|\mathcal{F}_0]| \geq \alpha) \leq 2 \exp\left(\frac{-\alpha^2}{2t}\right) \quad (118)$$

$$P(|\chi(G) - E[\chi(G)]| \geq \alpha) \leq 2 \exp\left(\frac{-\alpha^2}{2n}\right), \quad (119)$$

where we selected $t = n$ and noted that \mathcal{F}_0 leaves $\chi(G)$ unconstrained while \mathcal{F}_n defines it exactly. The result follows from choosing $\alpha = \lambda/\sqrt{n}$.

□

Exercise 8.3

a) Being an inner-product space $(\mathfrak{R}^n, \langle \cdot, \cdot \rangle)$ requires the dot-product $\langle \cdot, \cdot \rangle$ to fulfill (this is sufficient),

1. $\langle x, y \rangle = \langle y, x \rangle, \forall x, y, \in \mathfrak{R}^n,$
2. $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle, \forall x, y, z \in \mathfrak{R}^n, \alpha, \beta \in \mathfrak{R},$
3. $\langle x, x \rangle \geq 0, \forall x \in \mathfrak{R}^n$

Lets see what happens with our $k(x, z) = x^T A z$.

”if”. Need to show that if A is sym. pos. semdef., we have an inner-product space.

$$1. \langle x, y \rangle = k(x, y) = x^T A y = (x^T A y)^T = ((x^T A) y)^T \quad (120)$$

$$= y^T (x^T A)^T = y^T A^T x = y^T A x \quad (121)$$

$$= \langle y, x \rangle, \quad (122)$$

where we applied $x^T A y$ being scalar (its transpose is itself) and the symmetry of A .

$$2. \langle \alpha x + \beta y, z \rangle = (\alpha x + \beta y)^T A z \quad (123)$$

$$= (\alpha x^T + \beta y^T) A z \quad (124)$$

$$= \alpha x^T A z + \beta y^T A z \quad (125)$$

$$= \alpha k(x, z) + \beta k(y, z). \quad (126)$$

$$3. \langle x, x \rangle = x^T A x \geq 0, \quad (127)$$

due to pos. sem. def. of A being defined as $x^T A x \geq 0, \forall x \in \mathfrak{R}^n$.

”only” Asymmetry breaks condition 1. Being not positive semidefinite breaks condition 3. Hence, A must be symmetric and pos.sem.def.

b) Here we want to convert our kernel into a dot product form. Use the eigen-decomposition of A to get $V D V^T$. This leads to

$$x^T A z = x^T (V D V^T) z \quad (128)$$

$$= (x^T V) D (V^T z) \quad (129)$$

$$= (x^T V) \sqrt{D} \sqrt{D} (V^T z) \quad (130)$$

$$= (x^T V \sqrt{D}) (\sqrt{D} V^T z) \quad (131)$$

$$= \langle x^T V \sqrt{D}, (\sqrt{D} V^T z) \rangle \quad (132)$$

$$= \langle (x^T V \sqrt{D}), (z^T V \sqrt{D}) \rangle \quad (133)$$

$$, = \langle \phi(x), \phi(z) \rangle, \quad (134)$$

where the equivalences can be checked by examining matrix sizes of the results of the multiplications, and noting that D is symmetric.

□

Exercise 8.4

Since we already know that $(\mathfrak{R}^n, \langle \cdot, \cdot \rangle)$ is a Hilbert-space, if we select that, we only need to construct the proper mappings ϕ .

i) Choose $(\mathfrak{R}^{2^n}, \langle \cdot, \cdot \rangle)$ as the space. Define $\phi : Z \rightarrow 2^n$ s.t.,

$$\phi(A)_i = 1, \text{ if } S_i \subset A, 0 \text{ otherwise,}$$

where S_i is the i :th member of $\mathcal{P}(Z)$. That is, $\phi(x)$ has one dimension (index) for each possible subset. Now $k_1(A, B) = \langle \phi(A), \phi(B) \rangle$.

All binary vectors in the feature space are of the reqd. form.

ii) Choose $(\mathfrak{R}^n, \langle \cdot, \cdot \rangle)$, and suppose $P(A) = \sum_{a \in A} P(a)$. Now select

$$\phi(A)_i = \sqrt{P(a_i)}, \text{ if } a_i \in A, 0 \text{ otherwise.}$$

This results in $k_2(A, B) = \langle \phi(A), \phi(B) \rangle = P(A \cap B)$.

Loosely spoken, the suitable vectors are constrained by P 's nature as a distribution to be non-negative, and forced by $\sum a_i = 1$ to be inside some ball. The main point perhaps is that not all points of the feature space correspond to a dot-product between some two examples.

Exercise 8.5

Let $X = \mathfrak{R}^n, S = ((x_1, y_1), \dots, (x_m, y_m)), C > 0$,

$$V = \{\sum_i^m \alpha_i x_i | \alpha_i \in \mathfrak{R}\}, \text{ and}$$

$$R(w) = 1/2 \sum (w x_t - y_t)^2 + C/2 \|w\|_2^2.$$

Should show $w_* = \operatorname{argmin}(R) \in V$. Follow the given hint, choose arbitrary w from V , that is, arbitrary $\alpha \in R^m$. Now

$$R(w) = R(\sum_i \alpha_i x_i) = \frac{1}{2} \sum_t ((\sum_i \alpha_i x_i) x_t - y_t)^2 + \frac{C}{2} \|\sum_i \alpha_i x_i\|_2^2. \quad (135)$$

Denote with z some orthogonal addition. The first piece of the right hand side turns to

$$\begin{aligned} & \frac{1}{2} \sum ((\sum \alpha_i x_i + z) x_t - y_t)^2 \\ & \frac{1}{2} \sum ((\sum (\alpha_i x_i) x_t + z x_t - y_t)^2, \end{aligned}$$

noticing that $z x_t$ is zero due to the orthogonality, we see that we can not improve the square error term. However, the norm term (second on the right hand side) ends up as

$$\frac{C}{2} (\|w\|^2 + 2 \langle w, z \rangle + \|z\|^2),$$

that is, the norm can only increase ($\langle w, z \rangle = 0$).

Rant: the term on the right is called the *regularization term*, which can be loosely interpreted as an extra cost paid by the model complexity (here the norm). In this setting, a learning process can be interpreted as minimization of a cost function such as R . If we did regression and set C to zero, the minimization should result in standard least-squares linear regression with no concern for e.g. very high absolute values of the coefficients. Such an estimation procedure is easily thwarted by outlying examples (the square error is unbounded). Hence, regularization can be seen as both addressing overfitting and as a way to control the sensitivity of the learning process to errors in the data (which are not necessarily the same issue).

□