Practice Exam Machine Learning 2018
part 1 of 3

February 22, 2018

1. Which answer contains only *unsupervised* methods and tasks?
   A k-Means, Clustering, Density estimation ✓
   B Clustering, Linear regression, Generative modelling
   C Classification, Clustering, k-Means
   D k-NN, Density estimation, Clustering

2. In the book, Flach makes a distinction between *grouping* and *grading* models. Which statement is **false**?
   A Grouping models segment the feature space.
   B Grading models combine other classifiers, assigning a grade to each. ✓
   C Grading models can assign each element in the feature space a different prediction.
   D Grouping models can only assign a finite number of predictions.

3. We plot the ROC curve for a ranking classifier. What does the area under the curve estimate?
   A The probability of a ranking error ✓
   B The accuracy
   C The sum of squared errors
   D The probability of a misclassification

4. You want to search for a model in a discrete model space. Which search method is the **least** applicable?
   A Random search
   B Simulated annealing
   C Evolutionary methods
   D Gradient descent ✓

5. In bar charts, what do error bars represent?

    A Standard deviation
    B Standard error
    C A confidence interval
    D All are possible ✓

6. We can decompose the sample covariance matrix $\mathbf{S}$ into a transformation matrix as follows $\mathbf{S} = \mathbf{AA}^\mathsf{T}$. This allows us to transform normally distributed data into *standard* normally distributed data. However, the Principal Component Analysis doesn't use this decomposition, but the Singular Value Decomposition ($\mathbf{S} = \mathbf{UZU}^\mathsf{T}$). Why?
    A It's easier to compute.
    B There isn't always an $\mathbf{A}$ such that $\mathbf{S} = \mathbf{AA}^\mathsf{T}$.
    C It makes the loss surface more smooth.
    D It ensures the first axis has the highest eigenvalue. ✓

7. What is the relation between an ROC curve and a coverage matrix?
    A Normalizing the axes of the coverage matrix gives an ROC curve. ✓
    B Normalizing the axes of the ROC curve matrix gives a coverage matrix.
    C Dividing the values in the coverage matrix by the ranking error gives the coverage matrix.
    D The ROC curve is the transpose of the coverage matrix.

8. Which statement is **true**?
    A The average error of many models with high bias is low.
    B The average error of many models with high variance is low. ✓
    C A model with high bias has low variance
    D High bias is an indication of overfitting

Here we see the derivation of the gradient of the squared-error loss for linear regression. Which rules are applied in the indicated steps, to get from the line above it to the labeled line?

$$\frac{\partial \frac{1}{2} \sum_i \left(f(x_i) - y_i \right)^2}{\partial w} = \frac{1}{2} \frac{\partial \sum_i \left(x_i w + b - y_i \right)^2}{\partial w}$$

$$= \frac{1}{2} \sum_i \frac{\partial \left(x_i w + b - y_i \right)^2}{\partial w} \tag{1}$$

$$= \frac{1}{2} \sum_i \frac{\partial \left(x_i w + b - y_i \right)^2}{\partial \left(x_i w + b - y_i \right)} \frac{\partial \left(x_i w + b - y_i \right)}{\partial w} \tag{2}$$

$$= \sum_i \left(x_i w + b - y_i \right) \frac{\partial \left(x_i w + b - y_i \right)}{\partial w}$$

$$= \sum_i \left(x_i w + b - y_i \right) x_i \tag{3}$$

9. In line 1, we use the

    A Product rule **B** Chain rule **C** Sum rule ✓ **D** Exponent rule

10. In line 2, we use the

    A Product rule **B** Chain rule ✓ **C** Sum rule **D** Exponent rule

11. In line 3, the correct result is

    A $\sum_i \left(x_i^2 w + b - y_i \right)$
    B $\sum_i x_i \left(x_i w + b - y_i \right)$ ✓
    C $\sum_i \left(x_i w + b - y_i \right)$
    D $\sum_i \left(x_i w + b - y_i \right)^2$

We have the following training set:

	x_1	x_2	label
a	1	0	Ham
b	3	0	Ham
c	5	1	Spam
d	7	1	Spam
e	0	2	Ham
f	2	2	Spam
g	4	3	Spam
h	6	3	Ham
i	8	4	Spam

We use a linear classifier defined by

$$c(x_1, x_2) = \begin{cases} \text{Spam} & \text{if } x_1 + 0 \cdot x_2 - 2 > 0 \\ \text{Ham} & \text{otherwise.} \end{cases}$$

To answer the following questions, first draw the feature space, the classification boundary, and the coverage matrix.

12. If we turn c into a *ranking* classifier, how does it rank the points, from most "Ham" to most "Spam"?
    A e a f b g c h d i ✓
    B i g e h f c a b
    C d h b f i g c e a
    D e a i h g f d c b

13. How many ranking errors does the classifier make?
    A 2 **B** 4✓ **C** 6 **D** 8

14. What proportion of the coverage matrix is red?
    A $\frac{8}{20}$ **B** $\frac{3}{10}$ **C** $\frac{1}{5}$ ✓ **D** $\frac{1}{18}$