Practice Exam Machine Learning 2018

Full Exam

March 21, 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 $\mathbf{S} = \mathbf{A}\mathbf{A}^\top$. 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 ($\mathbf{S} = \mathbf{U}\mathbf{Z}\mathbf{U}^\top$). Why?

   A It's easier to compute.
   B There isn't always an $\mathbf{A}$ such that $\mathbf{S} = \mathbf{A}\mathbf{A}^\top$.
   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?

$$\begin{align}
\frac{\partial \frac{1}{2} \sum_i \left(f(x_i) - y_i\right)^2}{\partial w} &= \frac{1}{2} \frac{\partial \sum_i \left(x_i w + b - y_i\right)^2}{\partial w} \notag \\
&= \frac{1}{2} \sum_i \frac{\partial \left(x_i w + b - y_i\right)^2}{\partial w} \tag{1} \\
&= \frac{1}{2} \sum_i \frac{\partial \left(x_i w + b - y_i\right)^2}{\partial \left(x_i w + b - y_i\right)} \frac{\partial \left(x_i w + b - y_i\right)}{\partial w} \tag{2} \\
&= \sum_i \left(x_i w + b - y_i\right) \frac{\partial \left(x_i w + b - y_i\right)}{\partial w} \notag \\
&= \sum_i \left(x_i w + b - y_i\right) x_i \tag{3}
\end{align}$$

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** $\frac{1}{18}$

15. Consider the statement"The probability that the mean height of Dutch men is below 2 meters is high." Which is **true**?

    A A subjectivist would consider this an improper use of the term probability.
    B A true Bayesian would consider this an improper use of the term probability.
    C A true frequentist would consider this an improper use of the term probability.✓
    D In machine learning, we would consider this an improper use of the term probability.

16. Which statement is **false**?
    A Entropy is an expression of the uniformity of a probability distribution.
    B Entropy is an expectation of a codelength.
    C Entropy is expressed in bits.
    D The KL divergence is the least squares distance between two distributions.✓

17. Logistic regression fits a linear classifier by passing the result through $\langle 1 \rangle$, and applying a $\langle 2 \rangle$ loss. Fill in the blanks.

    A 1: a linear rectifier, 2: cross-entropy
    B 1: a linear rectifier, 2:least-squares
    C 1: a sigmoid function, 2: cross-entropy✓
    D 1: a sigmoid function, 2:least-squares

18. A convolutional layer reduces the number of weights (compared to a fully connected network). Which is **false**?

A It does this by setting the weights on connections to be equal.

B It does this by including *forget gates*. ✓

C It does this by not connecting every input to every node in the hidden layer.

D It does this by exploiting the locality encoded in the input.

19. It is difficult to find the maximum likelihood parameters for hidden variable models. Which is **false**?

    A If we marginalise out the hidden variable, we get a sum with a huge number of terms.

    B Optimizing the parameters directly by gradient descent often leads to mode collapse.

    C The EM algorithm often works but is not guaranteed to converge.✓

    D If the values of the hidden variables are given, the problem becomes tractable.

20. Which is **false**?

    A From a conditional distribution and a marginal distribution we can compute the joint distribution.

    B From the joint distribution we can always compute any conditional distribution.

    C From a conditional distribution we can always compute the joint distribution.✓

    D For independent random variables we can compute the joint distribution from their marginal distributions.

21. What is special about a *denoising* autoencoder?

    A It allows us to encode to a hidden layer bigger than the input. ✓

    B It directly optimizes the maximum likelihood.

    C It borrows ideas from the Expectation-Maximization algorithm.

    D It can be used for generative modeling.

22. I have an irrational fear of Lagrange multipliers. Can I derive the SVM algorithm without them?

    A Yes, but you can't use the kernel trick.✓

    B Yes, but you'll have to make do with an approximation.

    C Yes, but you can't use the result as the top layer of a neural network.

    D No.

23. I have a binary classification task. I build a Bayes classifier by fitting an MVN to each class. Which is **true**?

A The decision boundary is linear only if both MVNs have diagonal covariance matrices.
B The decision boundary is always linear.
C If the covariance matrices of both MVNs are the same, the decision boundary is linear.✓
D If the covariance matrices of both MVNs are the same, the decision is a (nonlinear) hyperbole.

24. We have a classification problem where the dataset is arranged on a straight line in a 3 dimensional space. The points are linearly separable along the line. Near the point of separation, some 3D Gaussian noise has been applied. Which classifier is most appropriate?

    A A logistic regression classifier. The noised points are *outliers* to which the LR classifier is robust.
    B A basic linear classifier. Others would pay too much attention to the points near the boundary, where the noise is.✓
    C A kernel SVM. A kernel could project these points into a higher dimensional space, where they are linearly separable.
    D A linear SVM. The maximum margin hyperplane criterion ensures that the linear nature of the dataset is taken into account.

Here are two distributions, p and q, on the members of a set $X = \{a, b, c, d, e\}$.

	p	q
a	$^2/_8$	$^6/_{16}$
b	$^2/_8$	$^1/_{16}$
c	$^2/_8$	$^5/_{16}$
d	$^1/_8$	$^3/_{16}$
e	$^1/_8$	$^1/_{16}$

25. What are their entropies?

    A $H(p) \approx 1.8, H(q) \approx 2.0$
    B $H(p) \approx 2.3, H(q) \approx 2.2$
    C $H(p) \approx 2.3, H(q) \approx 2.0$✓
    D $H(p) \approx 1.8, H(q) \approx 2.2$

26. What are their cross entropies?

A $H(p, q) \approx 2.6$,$H(q, p) \approx 2.3$✓

B $H(p, q) \approx 2.9$,$H(q, p) \approx 2.3$

C $H(p, q) \approx 2.6$,$H(q, p) \approx 2.5$

D $H(p, q) \approx 2.9$,$H(q, p) \approx 2.5$

We will use the backpropagation algorithm to find the derivative of the function

$$f(x) = \frac{\sin(1 + x^2)}{\cos(1 - x^2)}$$

with respect to x.

First, we break the function up into modules:

$$\begin{aligned} f &= \frac{a}{b} \\ a &= \sin c \\ b &= \cos d \\ c &= 1 + e \\ d &= 1 - e \\ e &= x^2 \end{aligned}$$

27. What should be in the place of the dots?

   A $a = \cos c$, $c = 1 + e$

   B $a = \sin c$, $c = 1 + e$✓

   C $a = \sin c$, $c = 1 - e$

   D $a = \cos c$, $c = 1 - e$

28. Work out the *local* derivatives. What is the correct expression for the gradient?

   A $\frac{\partial f}{\partial x} = \frac{\cos(c)2x}{b} - \frac{a\sin(d)2x}{b^2}$ ✓

   B $\frac{\partial f}{\partial x} = \frac{\cos(c)2x}{b} - \frac{\sin(d)2x}{ab^2}$

   C $\frac{\partial f}{\partial x} = \frac{\sin(c)2x}{b} - \frac{a\cos(d)2x}{b^2}$

   D $\frac{\partial f}{\partial x} = \frac{\sin(c)2x}{b} - \frac{\cos(d)2x}{ab^2}$

29. We want to find the minimum of the function $f(x, y) = \sin(x) + \tanh(x)$, subject to the constraint that $x^2 + y^2 = 1$ If we use the method of Langrange multipliers, what is our L-function?

   A $L(x, y, \lambda) = \sin(x) + \tanh(x) - \lambda x^2 - \lambda y^2 + \lambda$ ✓

   B $L(x, y, \lambda) = \cos(x) + \text{sech}^2(x) - \lambda x^2 - \lambda y^2 + \lambda$

   C $L(x, y, \lambda) = \sin(x) + \tanh(x) - \lambda 2x - \lambda 2y + \lambda$

   D $L(x, y, \lambda) = \cos(x) + \text{sech}^2(x) - \lambda 2x - \lambda 2y + \lambda$

30. Which does **not** describe the purpose of a regularizer?

A It simplifies selection of hyperparameters.✓
B It reduces overfitting.
C It functions as a definition of what constitutes a simple model.
D It biases the search algorithm towards simple models.

31. How do variational autoencoders avoid mode collapse?

A By training the "decoder" network through a discriminator.
B By using a regularizer to steer the network away from the data average.
C By feeding the discriminator network pairs of inputs.
D By learning the latent representation of an instance through an "encoder" network.✓

32. What is an *n-gram*?

A A short sequence of n tokens in sequential data.✓
B A latent representation of n dimensions.
C A way of storing data in a neural network.
D A batch of n instances.

33. What is the purpose of the word2vec method?

A It creates vector representations of words that embed semantics.✓
B It creates one-hot representations of words.
C It is an RNN model with increased ability to forget useless information.
D It ensures that Markov models can deal with unseen n-grams.

34. We train a generative model through plain gradient descent, using random examples from the data as targets, comparing these against random samples from the model, and backpropagating the error. After training, all samples from the model look like the average over the datasets. What is this phenomenon called?

A Multiple testing **B** Overfitting **C** Dropout **D** Mode collapse✓

35. What is meant by the "exploration vs. exploitation" tradeoff?

A That hyperparameter selection (exploration) uses computational resources that can also used in training the model (exploitation).
B That an online algorithm needs to balance optimization of its expected reward with exploring to lean more about its environment.✓
C That an insufficiently thoroughly trained model may be biased against minorities.
D Balancing the loss function with the regularization terms in matrix factorization.

36. What is the *Markov assumption*?

A We can apply backpropagation to neural networks by unrolling them.
B The probability of a word does not depend on the words preceding it.
C The operation of an LSTM cell depends only on its predecessors through two inputs.
D A word is conditionally dependent only on a finite number of words preceding it.✓

37. In recommender systems, what is *implicit feedback*?

A Associations between users and items assumed from user behavior.✓
B Ratings given by a single “like” button rather than a more fine-grained system.
C Recommendations that take the temporal structure of the data into account.
D Recommendations derived from manually crafted item features rather than learned ones.

38. In PCA, Which regularizer can we use to create sparse representations?

A L1 ✓ **B** L2 **C** Dropout **D** All three

The EM and VAE algorithms are both based on the following decomposition.

$$\begin{aligned}
L(q, \theta) + \text{KL}(q, p) &= \mathbb{E}_q \ln \frac{p(x, z \mid \theta)}{q(z)} - \mathbb{E}_q \ln \frac{p(z \mid x, \theta)}{q(z)} \\
&= \mathbb{E}_q \ln \frac{p(x, z \mid \theta)}{q(z)} - \mathbb{E}_q \ln \frac{p(z \mid x, \theta)}{q(z)} \\
&= \mathbb{E}_q \ln p(x, z \mid \theta) - \mathbb{E}_q \ln q(z) - \mathbb{E}_q \ln p(z \mid x, \theta) + \mathbb{E}_q \ln q(z) \\
&= \mathbb{E}_q \ln p(x, z \mid \theta) - \mathbb{E}_q \ln p(z \mid x, \theta) \\
&= \mathbb{E}_q \ln \frac{p(x, z \mid \theta)}{p(z \mid x, \theta)} = \mathbb{E}_q \ln \frac{p(z \mid x, \theta) p(x \mid \theta)}{p(z \mid x, \theta)} \\
&= \mathbb{E}_q \ln p(x \mid \theta) = \ln p(x \mid \theta)
\end{aligned}$$

39. What should be in place of the blanks?

A $p(x \mid z, \theta)$
B $q(x \mid z, \theta)$
C $p(z \mid x, \theta)$ ✓
D $q(z \mid x, \theta)$

40. Why does this decomposition help us?

A It eliminates q, which we cannot compute.
B It shows that $L(q, \theta)$ is a lowerbound on the quantity we want to maximize.✓
C It eliminates expectations, which we cannot compute.
D It shows that $p(\mathbf{x} \mid \theta)$ is a lowerbound on the KL divergence.