MID-TERM EXAM ML.

① Overfitting occurs when the generalization error worsens even when the training error keeps being reduced by the learning algorithm.

- True:

Overfitting is the increment of test error when a certain complexity level is attained.

② We can use regression algorithms to solve classification problems.

- True

A linear model in n dimensions can solve any classification problem.

③ Recall tells us the proportion of the positive class with respect to all data predicted positive.

- False

$$\text{Recall} = \frac{TP}{\text{Real Positives}} = \frac{TP}{TP+FN}$$

④ Features encoded by hashing show a close to uniform distribution.

- True

Distribution of hashed data tends to be uniform.

⑤ The (i,j)-th element of the confusion matrix counts the amount of samples from class i that take value j

- False

Confusion matrix has $\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}$

⑥ In general, removing a feature with a lot of missing data is a bad idea.

- False

Deletion when a lot of missing data is present on a sample is not a bad idea.

Escaneado con CamScanner

⑦ L1 regularization may help in selecting features when combined with a linear model.

• True

L1 regularization, min $\|\alpha\|_1$ gives sparcity, since implicitly forces most of the parameters of a linear model $\hat{y} = \sum \alpha_i f_i(x)$ to be 0 ⇒ implicit feature selection.

⑧ The natural loss for a classification problem is the 0-1 loss.

• True.

Ideal loss function for classification: 0-1 loss.

⑨ In a marketing campaign launching based on machine learning churn (positive class) prediction, if we want to maximize the amount of recovered clients we want to maximize the recall value classifier.

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{\text{True P}}{\text{Predicted P}}$$

we want that all the predicted positives correspond to the true positives.

• True.

⑩ In unbalanced datasets the use of accuracy is not informative of the performance of the classifier

• True

Unbalanced datasets: the value of correctly predicting elements from different classes is different → we need different metrics: confusion matrix.

⑪ Hashing is an encoding method that allows to work with an indefinite set of categories.

• True

Hashing doesn't need to prepare a dictionaire or structure, it works with indefinite sets of categories

⑫ In a regression problem we know that we have samples that come from the true generating function without noise. Disregarding the number of samples, I will select a model with the same complexity as the true generating function.

• False.

Independently of the data generation process generating we have to watch the data complexity, not the model complexity.

Escaneado con CamScanner

(13) LOO maximizes the amount of data used for training purposes while still being able to produce a good estimation of the validation error.

- True

LOO every point used as train/test ⇒ we get good estimation of validation error, instead of an a unique random variable.

(14) An orange in 100-dimensional space has more pulp than peel.

- False.

In high dimensions { Mass of multivariate Gaussian distribution is not near the mean but in an increasingly distant shell around it

(15) A copy C is a model that aims at copying the decision boundary of another original O. In order to do so, we generate random samples and label them according to the original prediction on those samples. In this setting, overfitting is not a concern.

- True. { Data augmentation: $\hat{x} = x + \eta$ ⇒ DROP OUT TECHNIQUE ↗ we cancel the attributes at random when training, curing overfitting.

(16) Learning means that we can directly optimize the generalization error.

- False.

Learning consists of finding a model such that $E_{out} \to 0$ (minimizes the out of sample/ generalization error). In order to do so:

1) $E_{in} \to 0$ (min E_{in})

2) $E_{in} \approx E_{out}$ (Hoeffding).

(17) A good way of encoding categorical variables is ordinal encoding: assigning a numerical value to each categorical value.

- False

Categorical { Dummy variables / Hashing

Numerical values introduce a different metrics

(18) A classifier in the hypothesis space is characterized by a point.

- TRUE

A classifier from a model class is characterized by the corresponding parameters represented by points.

Escaneado con CamScanner

(19) In order to avoid overfitting I will select the model that has the minimum training error when we change the complexity of the classifier.

· False

Training error keeps smaller with complexity increase, we seek the minimum of the testing error! (that may increase after a certain value of complexity.)

Correcting it by cross-validation, regularization or ensemble techniques.

(20) Cross-validation is a simulation method for avoiding overfitting.

· True

(21) Training Error to be zero is a necessary condition for a learning problem to be feasible.

· False.

Feasibility of a learning problem occurs when $\begin{cases} E_{in} \approx E_{out} \\ E_{in} \approx 0 \end{cases}$

(22) We can reduce overfitting by reducing the complexity of the classifier.

· True.

Combating OVERFITTING: min complexity (REGULARIZATION).

(23) Unsupervised learning aims at finding a decision boundary when data is not labled.

· False

Unsupervised learning: given $\{x_i\}_{i=1}^{N}$ seeks to find their structure: density estimation, clustering...

(24) VC dimension is a measure of the complexity of a classifier.

· True, it is the max # of points it can shatter.

(25) Selecting the value of k of k-nearest neighbours stands for changing the hypothesis space:

· TRUE: we are setting the parameters of the learning model ∈ hypothesis space.

Escaneado con CamScanner

(26) We can avoid overfitting by increasing the # of samples.

- True, $E_{out} \leq E_{in} + O\left(\sqrt{\frac{c}{N}}\right)$

(27) L2 regularization may help in selecting features when combined with a linear model.

- False, $\min \|x\|_2$ not necessarily gives sparcity

(28) The perfect operational point has TPR = 1, FPR = 0.

$TPR = \frac{TP}{TP+FN}$ = recall

$FPR = 1 - \text{specificity} = \frac{FP}{TN+FP}$

For the perfect operational, FN = 0 = FP, then TPR = 1, FPR = 0

- True.

(29) When we normalize training and test data we find the normalization values for each of the two sets independently

- False.

We scale both by the same parameters, training and test set

(30) The role of regularization is to model complexity of the classifier.

- True:

REGULARIZATION: modelization of the model's complexity

(31) Specificity tells us the proportion of the positive class with respect to all data predicted positive.

$\text{Specificity} = \frac{TN}{TN+FP}$

- False: negative real class with respect to all ~~predicted~~ negative real.

(32) In general, if the source generating data is a 100th-order polynomial, I would use a 100th order polynomial as a model.

- False.

We have to match the data complexity, not the real model's.

Escaneado con CamScanner

(33) VC dimension counts the number of points that can be shattered by any function of the hypothesis space.

- True, by definition

VC DIMENSION: max # of points a classifier can shatter.

(34) Negative predictive value is the equivalent to the specificity for the negative class.

Specificity: $\frac{TN}{TN+FP}$

Sensitivity /recall $= \frac{TP}{TP+FN}$

Negative predictive value $= \frac{TN}{TN+FN}$

Positive predicted value/Precision $= \frac{TP}{TP+FP}$

$\frac{\text{True}}{\text{Predicted}}$

$\frac{\text{True}}{\text{Real}}$

- False

(35) In a cross-validation process all samples are used for testing.

- True.

CROSS-VALIDATION TECHNIQUES: splits the data set in different sets disjoints so we can train the model several times and test it each time with the left out set, so all the data set samples are eventually used for testing.

Escaneado con CamScanner

(36) Precision tells us the proportion of the positive class with respect to all data predicted positive:

$$\text{Precision} = \frac{\text{True Positive}}{\text{Predicted Positive}} = \frac{TP}{TP+FP}$$

· True

(37) In train-test splitting we want the number of samples of test to be the largest possible.

· TRUE $\boxed{\sim}$

(~~the~~ aim a balance between the train and test sets that allows us to satisfying train ~~for~~ the model (around 70% training))

(38) In a time-series it is reasonable to replace missing data by the mean of the series:

· False.

Time-series don't usually have simple distribution (they in fact can have some monotony)

(39) Hashing is an encoding method that preserves metrics of the original space.

· False

Hashing ⇒ original metric disappears.
In fact, the hashed distribution tends to be uniform.

(40) The use of raw data is unadvisable because of its unknown discriminative power.

· FALSE

Raw data disadvantages { highly redundant + large dimensional spaces
Unknown discriminability.

Is not unadvisable.

(41) Curse of dimensionality means that we are in front of a very difficult problem.

CURSE OF DIMENSIONALITY: many algorithms that work fine in low dimensions become intractable when the input is high dimensional.

ML → generalizing correctly becomes exponentially harder as the dimensionality (# of features) of the examples grows; because a fixed-size training set covers a dwindling fraction of the input space.

· FALSE

Escaneado con CamScanner

(42) In a rare disease problem where we do not want to lose any infected patient (positive class) we want to maximize the recall value of the classifier

$$\text{Recall} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{Real Positives}} = \frac{TP}{TP+FN}$$

• True

(43) $E_{out} \geq E_{in}$

• True.

E_{in}: insample error = freq. of hypothesis getting it wrong : TRAINING ERROR
E_{out}: out of sample error = expected error : TESTING ERROR.

(44) Regarding the two learning curves plots, the phenomenon of overfitting can only be observed when we check the training and test error for a fixed number of samples when the complexity varies.

• False.

We can also see overfitting with different fixed complexities and changing number of samples. Overfitting can be observed when the BIAS (value to which train and test error tend when the # of samples increases) gets greater when we increment the complexity for different plots
< decrease of !

(45) F1-score is a good metric for unbalanced datasets.

• True:

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = 2 \frac{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} = \frac{2\,TP}{TP+FN+TP+FP}$$

(46) In general, the more features we have, the better for classification purposes.

• False:

More features without information just add noise.

Escaneado con CamScanner

(47) Small complexity models are preferable in general because they avoid overfitting.

- True

Overfitting occurs when the test error increases while train error keeps decreasing as complexity increases.

(48) Given a training example, setting the value of one of the features of that sample to fixed value helps avoiding overfitting.

- True $\boxed{N}$

(49) In general, if the source generating data is 1st-order polynomial, I would use a 1st order polynomial as a model.

- False.

We match the data complexity, not the model's.

(50) Intuitively, we could describe an orange 100-dimension as a star-like fruit.

- True

(51) Adding noise to input data helps avoiding overfitting.

- True, since $E_{out} \leq E_{in} + O\left(\sqrt{\frac{C}{N}}\right)$

$\uparrow$ data $\Rightarrow$ $\downarrow$ overfitting

$\downarrow$ complexity $\Rightarrow$ $\downarrow$ overfitting

(52) $E_{out} \leq E_{in} + O\left(\frac{N}{C}\right)$

- False

(53) In Hoeffding's bound, if we take the tolerance value, epsilon, inside the probability term, we need to double the number of samples for the probability value to stay the same.

Hoeffdings bound: $P(|E_{out} - E_{in}| > \varepsilon) \leq 2e^{-N\varepsilon^2/2}$

$\varepsilon \to \varepsilon/2$: $P(|E_{out} - E_{in}| > \frac{\varepsilon}{2}) \leq 2e^{N\frac{\varepsilon^2}{4}/2} = 2e^{N\varepsilon^2/8}$

For the probab to stay the same we would need $N \to 4N$.

Escaneado con CamScanner

(54) When modelling the classification problem we strictly consider the cost/loss function that models the "irritation" we feel when a sample is misclassified.

- FALSE

Cost function: quantifies the effect of misclassifying a sample.

(55) In a regularized optimization problem increasing the weight of the term of the regularization decreases the complexity of the solution model.

- TRUE

We seek to minimize the weights of the regularization if we want complex. ($\min \|x\|_1, \|x\|_2$) By increasing it, we decrease complexity

(56) The generalization error following the training error is a necessary condition for any learning problem to be feasible.

- True, we need $\begin{cases} E_{in} \approx E_{out} \\ E_{out} \to 0 \end{cases}$

(57) The ROC curve requires changing a parameter of the classifier for its plot.

- True.

!

(58) Two different hypothesis spaces can intersect (2 different models may display exactly the same boundary)

- True.

(59) We can avoid overfitting by means of using ensemble methods.

- True.

(60) The three components when defining a ML model are: ① deciding the hypothesis space, ② selecting the loss function (modelling the problem) and ③ finding the model parameters that best fits the data.

- True.

Escaneado con CamScanner

TEST ML 1:

① When using ML techniques, the accuracy reported in the training set is a good indicator of the performance we will obtain when applying the method in practice:

- True.

④ Model selection is not the method to use for setting hyper-parameters in a model, for ex deciding the value of the number of neighbors.

- False.

MODEL SELECTION: deciding the model (with its parameters) we may use in the learning process.

⑤ If a previously selected model achieves worse performance than another model on the test set we will change the method to the best performant one.

- False! We are not supposed to contract the results from the test set. If we do so, this may actually be a validation set.

Escaneado con CamScanner