3.1 & 3.2

1. False - The ultimate goal of ML is to have low error on the test data.
2. In ML, a test set …

* should adhere to
* can reuse some samples from the train set
* is usually assumed to be sampled from the same distribution as the train set
* must be smaller than the train set

1. Consider the following scenario.

We want to train a ML model many times to find hyperparameters that yield good generalization performance. We are careful to ensure that every trial uses random splits that respect . What is the problem with this setup (if any)?

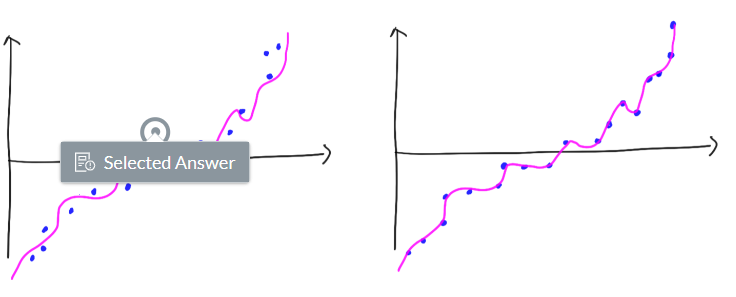
* The training and test sets should have a few samples in common so that we can calibrate our model
* Random splits will not yield reliable generalization estimates because they are not truly random
* Using the test set to select hyperparameters must be avoided to ensure a good estimate of generalization

1. Consider the following ML models applied to training data.



1. The figures below show a model that is over fitting. One panel shows the model on the training data, the other shows the model on the test data.

**Click** on the panel that shows the model on the test data.



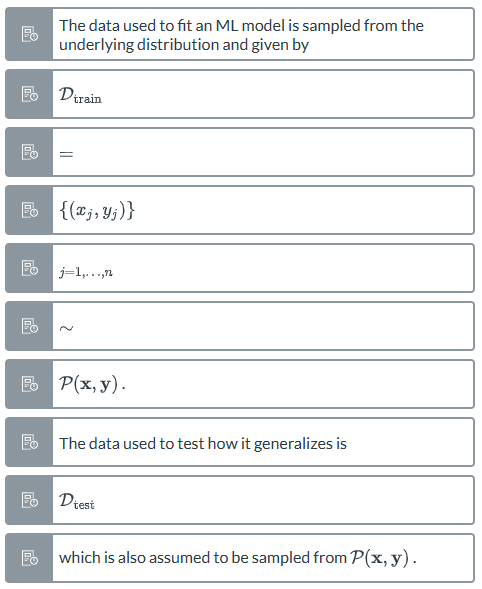
1. **Fill in** the missing parts of the following statement to make it correct.

An **underfitted** model usually lacks the **expressiveness** to capture the underlying structure of the data. An **overfitted** model has more expressiveness than is necessary to fit the **underlying** data. During **training**, that expressiveness may unintentionally capture some of the **natural variance** in the training set resulting in a **poor fit** when applied to unseen data.

* If a model has bad performance on the test set it is overfitted.
* An extreme case of overfitting is to memorize the training data.
* If a model has bad performance on the test set it is underfitted.
* If a model has very good performance on the train set it is overfitted.
* If performance is better on the test set than on the training set, the model is overfitted.

1. **Arrange** the following to form a correct statement.

Hint: use fullscreen mode if you have trouble manipulating the terms.



1. False - In practice, machine learning only works well if the *i.i.d. assumption* is not violated.
2. **Select** the following examples that **do not** violate the *i.i.d. assumption.*

* A collection of football match results from an entire season up to the championship divided randomly into training and test sets.
* Stock values of tech companies (AAPL, MSFT, GOOG, AMZN, FB) in January (training set) and February (test set)
* A collection of images of street signs from Switzerland (training set) and Sweden (test set)
* A collection of all the photos on Facebook from users in Sweden randomly split into training and test sets.

1. **Select** the best statement from below.

For a fixed training set, if we were to sequentially add parameters to give more expressiveness to the model, we are more likely to observe:

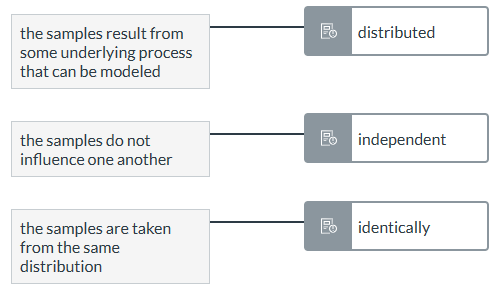
* an increase in training error
* a decrease in the training error
* no change in training error
* a decreasing difference between the training and test errors

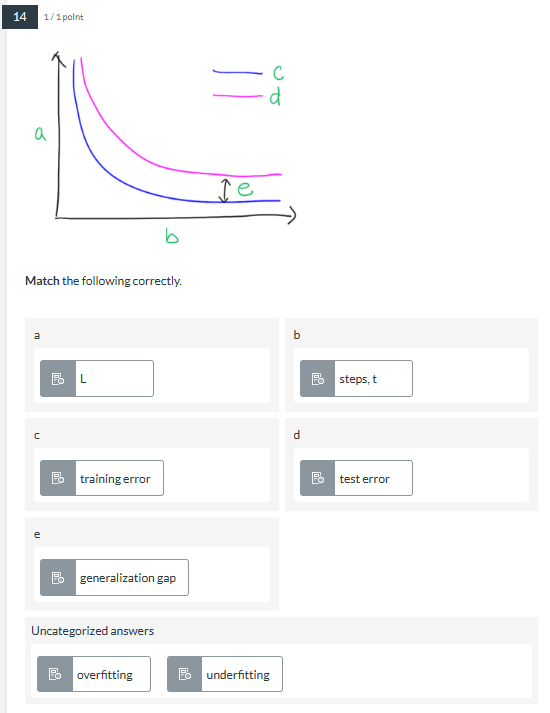
1. **Select** the best statement from below.

For a model with fixed parameters, if we were to sequentially add more training data and retrain the model, we are more likely to observe:

* a decrease in the training error
* a decreasing difference between the training and test errors
* an increasing difference between train and test errors

1. **Match** the following terms from the i.i.d. assumption to their meaning



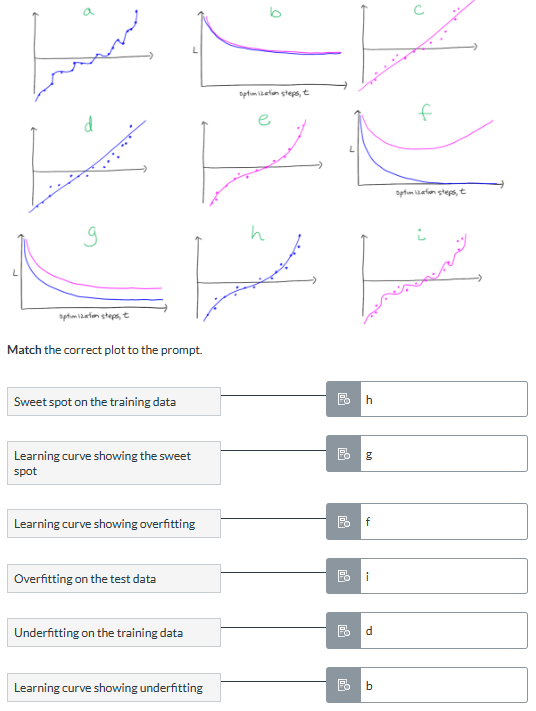


1. **Mark** the following that are correct.

* If the test error of the model is low, it probably generalizes well.
* A model with very high training error is unlikely to generalize well.
* A model with very low training error cannot generalize well.

1. **Fill in** the following statement correctly.

In DD1420, **error** is the difference between a prediction and its **true value**. **Loss** is a choice we make in our modeling to measure **error**. Because of their similarity, these terms are sometimes used interchangeably. The **train error** refers to the average loss on the training set. The **test error** refers to the average loss on the test set.



1. **Mark** the correct statements below.

* Length of the training has an effect on under/overfitting.
* Usually it is a good strategy to halt training once training error is low enough.
* Usually it is a good strategy to stop training once validation error starts increasing.

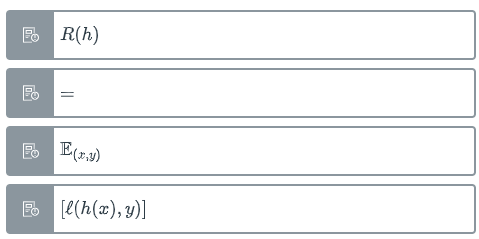
1. **Fill in** the missing words to correctly complete the following statement.

In DD1420 we introduce three types of learning curves. We mainly focus on the curve that plots the **loss** as a function of **optimization steps**. Another type plots it as a function of **dataset size** where the **generalization gap** shrinks as **data is added**. In this version, the model is fixed and assumed to be fully optimized. A third type plots model **capacity** on the x-axis, showing the effect of adding expressivity to the model. In this version, the dataset is fixed and the model is assumed to be fully optimized.

3.3

1. true - In learning theory, we assume that there exists a (probably unknowable) data generating process that the data we work with is i.i.d. sampled from. Knowledge of the data generating process is something akin to an omniscient view of where the data came from.
2. **Arrange** the following terms into the expression for risk.

Hint: use fullscreen mode if you have trouble manipulating the terms.



1. **Mark** all of the following statements that are correct.

* Minimizing the risk means good generalization
* The risk of hypothesis ℎ is the expectation of the loss/error over the data generating distribution.
* The risk is akin to an omniscient view of the average loss.
* Risk is the expectation of the hypothesis' mistakes over all possible x and y

1. Which of the following represent a valid hypothesis ℎ?

**Mark** all that apply.

* A hyperplane defined by b=2 and w=0.25
* =0
* Soft-margin SVMs with C=1.0
* A logistic regression model with initial parameters set to zero.
* A linear regression model with =[0 0.4 0.2]

1. **Mark** all statements below that are correct.

* A linear SVM defined over with =[1 0.2 1] defines hypothesis class
* The set of all hyperspheres around a point x with radius r defines a hypothesis class
* The set of all linear SVMs and polynomial kernel SVMs defines a hypothesis class
* A hypothesis space contains all valid hypotheses that can be defined by a model

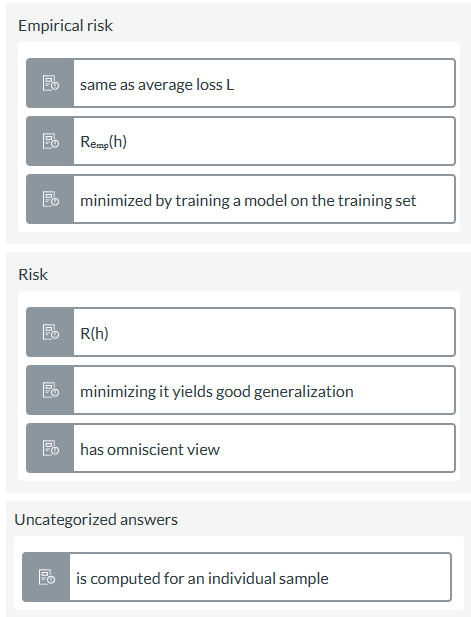
1. True - Risk is difficult to compute in practice because the distribution P is unknowable (usually).
2. **Mark** the correct statements.

* We usually minimize the risk using optimization techniques
* A randomly initialized logistic regression model is not a hypothesis
* If we know that R(ℎ1)≥R(ℎ2), we should use model ℎ2
* Risk of a training sample is usually lower than a test sample
* Risk is defined for a (specific) model

1. False - Risk is defined for the test set.
2. True - In the limit , and are the same.
3. **Fill in** the following statement correctly.

The **empirical risk** is the same thing as the **average loss** L. Because we can't know how an ML model will work on the **true distribution**, we estimate it **empirically** using the **training data.**

1. False - Our ultimate aim is to minimize the empirical risk , and the risk is just a proxy.
2. empirical risk



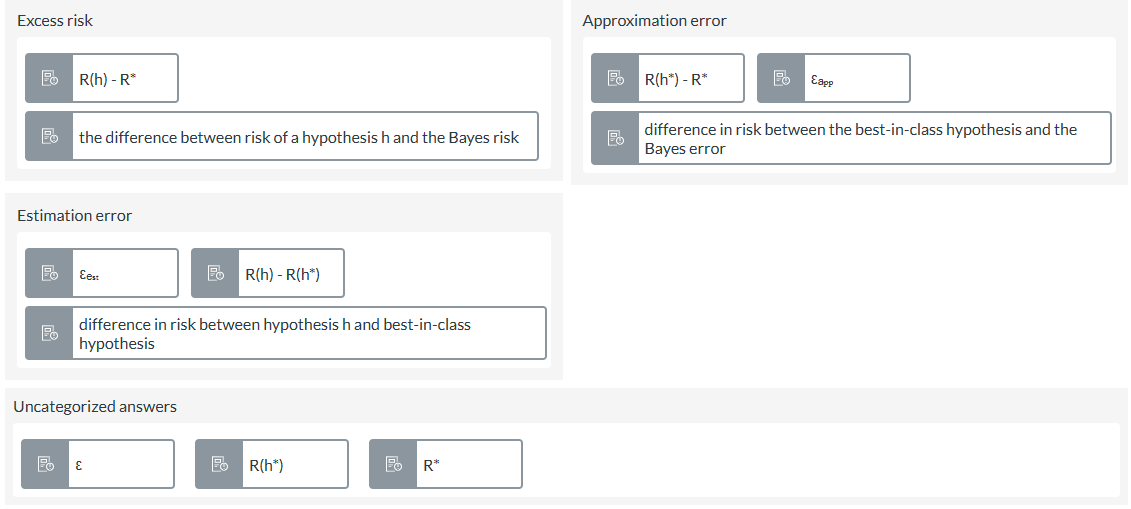
1. **Fill in** the following statement so it is correct.

The **Bayes error, R\* = infₕ R(h),** is the **lowest** possible risk you can get over all possible **hypotheses.**

1. According to the notation conventions from the Lecture Notes

**Mark** all of the correct statements.

1. **Match** the following to the correct category.

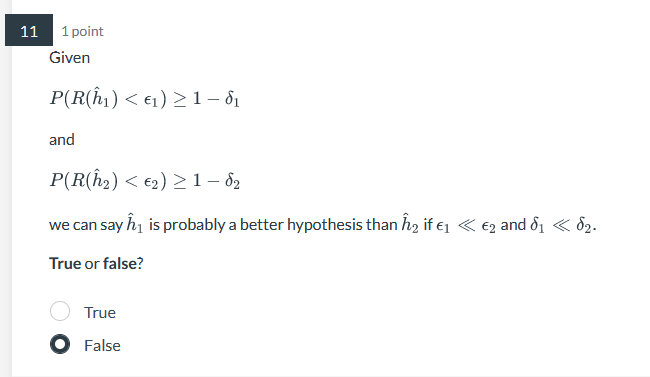


1. Assume you are not allowed to change your training algorithm or your training data. What can you do to improve your best-in-class hypothesis ?

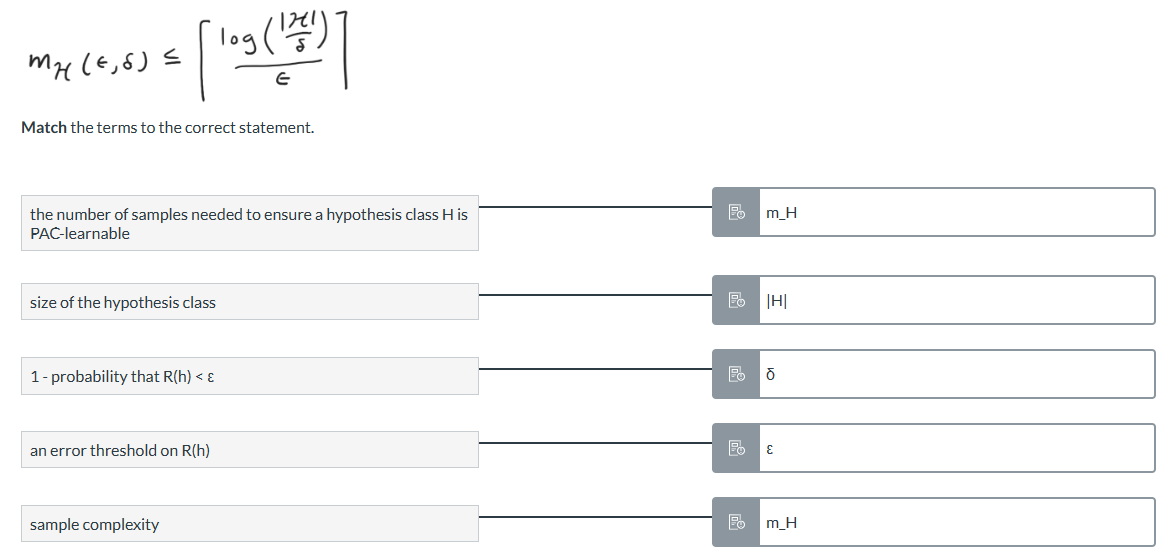
**Mark** the following statements that are correct.

* Increase the risk of the Bayes optimal hypothesis
* Reduce the hypothesis class to include fewer model types
* Expand the hypothesis class to include more model types
* Reduce the risk of the Bayes optimal hypothesis

1. true - Give and we can say if and .



1. Consider the expression for *sample complexity*



1. If an H is PAC-learnable, **mark** the correct statements.

* We can probably learn an that can approximate the Bayes optimal hypothesis.
* There will be some distribution where the learned hypothesis will never be able to approximate the best-in-class hypothesis.
* We can probably learn an that can approximate the best-in-class hypothesis

1. **Fill in** the missing terms to complete the statement correctly.

A **hypothesis class H** is PAC-learnable with sample complexity m\_H if some algorithm A exists that can produce **hypotheses** such that for all distributions, there is a probability of **1-δ** that the risk of the **learned hypotheses** are within **ε** of the **best-in-class hypothesis h\*** (if the size of the dataset is equal or larger to **m\_H**).

1. The fundamental theorem of PAC learning tells us...

**Mark** all the apply

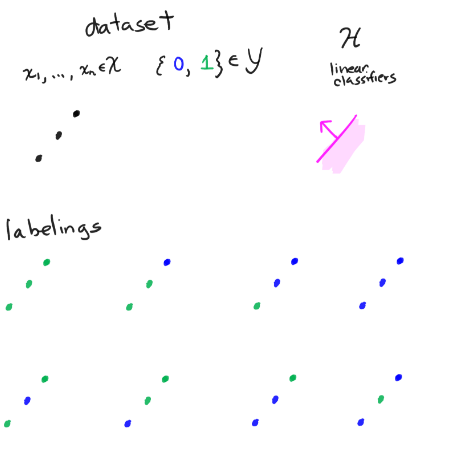
* that a more expressive hypothesis class may reduce the ability to generalize
* the worst case difference between the training error and the risk
* that increasing the size of the training data may improve generalization
* the conditions necessary to learn a hypothesis that generalizes well

1. When training a ML model in practice, it is not common to check the bounds given by the fundamental theorem of PAC learning because...

**Mark** all that apply

* The VC dimension is difficult to compute for all but the most simple datasets and hypothesis classes
* The fundamental theorem gives an upper bound on the worst-case-scenario, but doesn't say much about what we should *typically* expect from our trained models
* In practice, we never know which hypotheses belong to our hypothesis class H
* The probability is difficult to compute

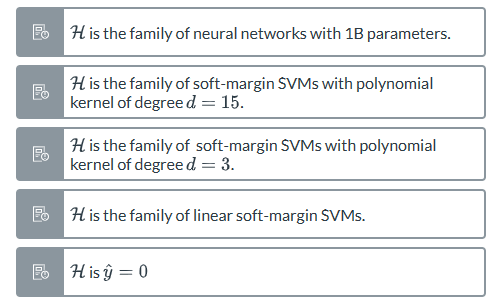
1. true - The Vapnik-Chervonekis (VC) dimension measures the complexity of a hypothesis class H, or how expressive it is.
2. Consider the following dataset containing 3 points with binary labels {0,1}. Can the set of linear classifiers defined in this space H *shatter* this dataset?



* No, it cannot shatter because it fails in the bottom right case
* Yes
* No, it cannot shatter because it fails for the case in row 1, column 3
* No, it cannot shatter because it fails in the bottom left case

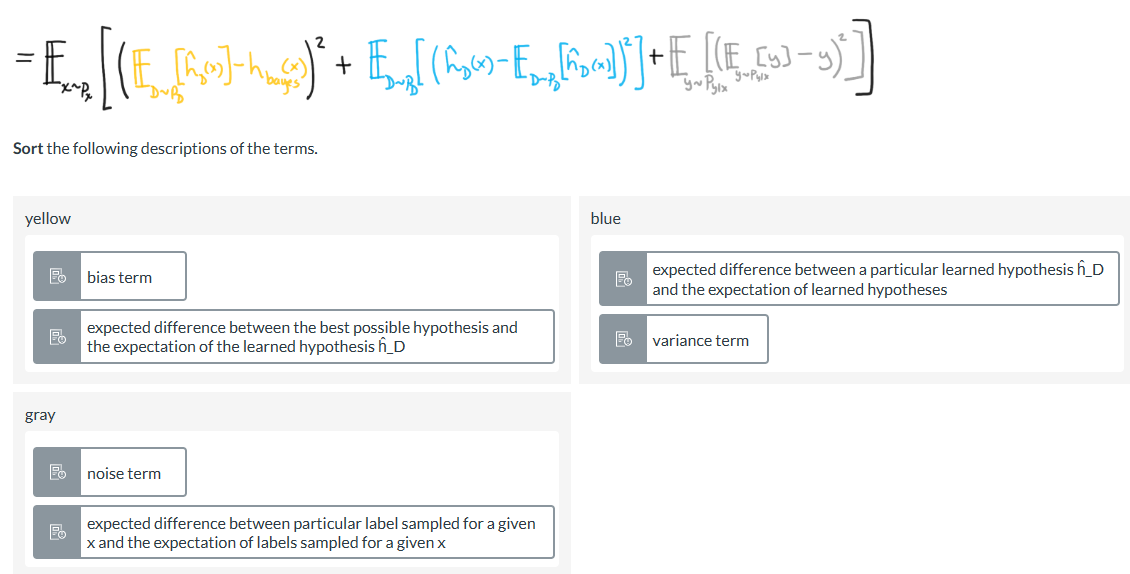
1. **Sort** the following in terms of model complexity (or expressivity) on a dataset with 10,000 training examples.

most expressive



least expressive

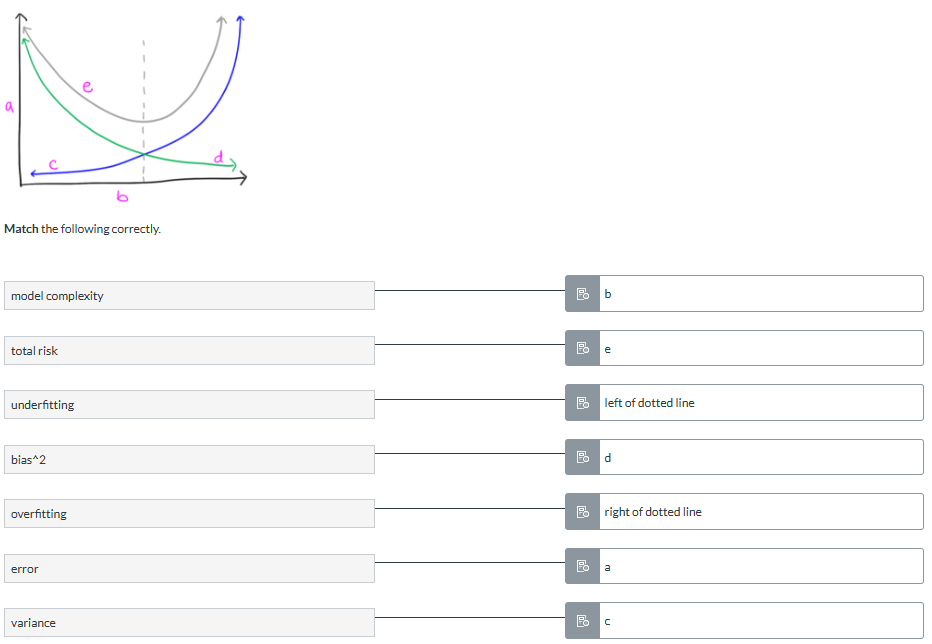
1. Consider the following decomposition of the risk



1. **Mark** all of the correct statements about bias and variance below.

* An expressive model will always have high variance.
* Variance tells us how much our model would change if we learned it on different datasets sampled from .
* A restrictive model will have high bias compared to
* Bias and variance are affected by our modeling decisions.

1. Consider the following diagram



29. **Fill in** the following statement correctly.

The **no free lunch theorem** states that there is **no single learning algorithm** that can yield hypotheses with **good generalization** for **all distributions**

30. **Mark** all the following statements that are correct.

When would the empirical risk minimization principle (ERM) likely work in practice?

* The training set is larger than the test set, but they come from the same distribution.
* A training set of street view images was collected in New York and a test set was collected in Stockholm
* A training set sampled from a Poisson distribution was generated on a computer in Tokyo. A test set was generated using the same code and parameters but using a different random seed on a computer in Stockholm.
* A large training set and test set of equal size that come from the same distribution
* The training set of customer data was collected before the the pandemic, and the test set was collected during lockdown.

Quiz 3.4

**1**. Given that a classifier has 15% accuracy, mark the true statement:

✖️The classifier is not performing well

✅There is not enough information to tell if the classifier is performing well or not

✖️The classifier is performing well

**2.** We have a binary classification task with 100 samples. Our classifier has a recall of 10%. Please fill in all possible values that the precision could take:

✅100%

✅50%

✖️0%

✅10%

**3**.F1 is used more often thanFβ since it is easier to use and compare.

✖️Not enough information

✖️False

✅True

**4.**A linear ROC Curve (y = x) means that the model is learning useful things.

✖️Not enough information

✅False

✖️True

**5.**The “recall” metric tells us what proportion of the actual positive cases that our classifier was able to predict as positive.

✖️Not enough information

✖️False

✅True

**6.**A lower Fi is generally considered better.

✖️Not enough information

✅False

✖️True

**7.**All ROC Curves pass through (1, 1).

✖️Not enough information

✖️False

✅True

**8.** All ROC Curves pass through the origin (the point (0, 0)).

✖️Not enough information

✖️False

✅True

**9.** The “precision” metric tells us what proportion of the predicted positive cases that actually have the positive condition.

✖️Not enough information

✖️False

✅True

**10.**Given that a classifier has 100% accuracy on the test set, mark the true statement:

✖️The classifier is not performing well

✖️There is not enough information to tell if the classifier is performing well or not

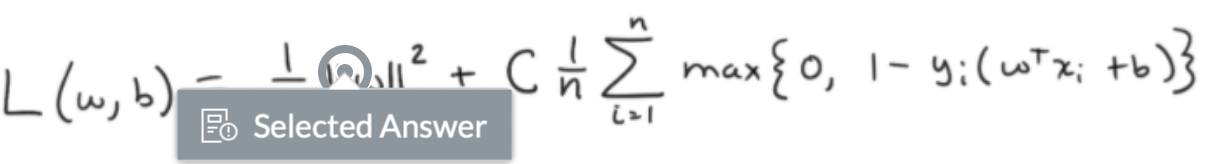
✅The classifier is performing well

Quiz 3.5 and 3.6

**1. Fill in** the following statement correctly.

From learning theory, we saw that the fundamental theorem of PAC learning and the bias-variance tradeoff imply that **overly expressive models** may end up **performing poorly** in practice. A simple and effective way of dealing with this is to add a

**regularization** term to the **objective function** which **limits** the numeric range of the model parameters.

**2.Click** on the regularization term in the following loss for a soft-margin linear SVM.

**3.** Suppose you are using logistic regression to perform classification. The loss function is given by  Hint: use fullscreen mode if you have trouble manipulating the terms.

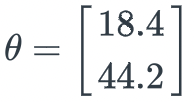
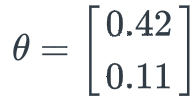
which we will simplify as **Complete** the following statement correctly.

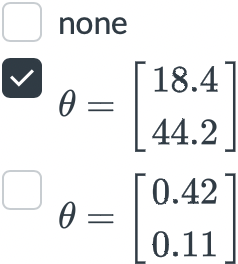
To apply **L1** (or lasso) regression, the equation becomes **L(θ) = 1/n Σ 𝓁(h(xᵢ), yᵢ) +αΣ |θ| .**

**4.** Suppose you are using logistic regression to perform classification. The loss function is given bywhich we will simplify as  **Complete** the following statement correctly.

To apply **L2** (or ridge) regression, the equation becomes **L(θ) = 1/n Σ 𝓁(h(xᵢ), yᵢ) + α Σ ||θ||²** .

**5**. Suppose you are doing hyperparameter selection while training a logistic regression model on your dataset. The first time you tried with regularization parameter α=0 and the second time you tried with α=1. One time, you got the model parameters

and the other time you got the model parametersWhich do you think corresponds to α=0.



**6.** Suppose you are training a logistic regression model. Which of the following statements are true?

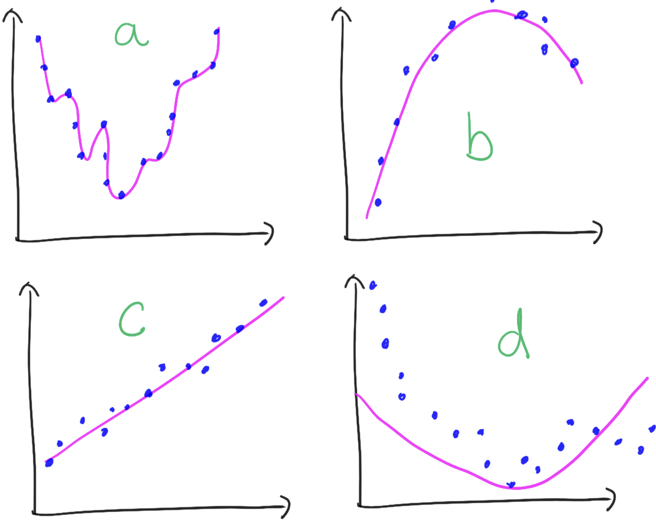
**Check** all that apply.

✖️Introducing regularization to the model always results in equal or better performance on unseen examples (outside the training set)

✖️Introducing regularization to the model always results in equal or better performance on the training set.

✖️Introducing regularization to the model will make the model underfit

✅Introducing regularization to the model limits its expressivity

**7**. Consider the following models that were fitted to training data.

Which of the models is *most* likely to benefit from **regularization**?

**Select** the correct answer.

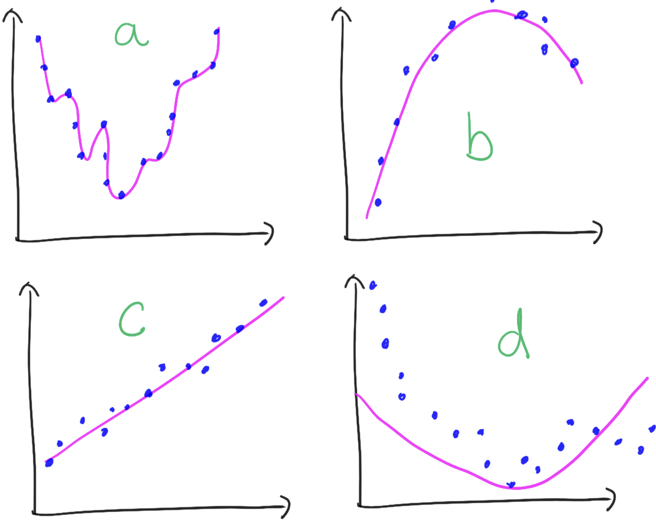
✖️d

✅a

✖️c

✖️b

**8.** Consider the following models that were fitted to training data.



Which of the models is ***least* likely** to benefit from **regularization**?

**Select** the correct answer.

✅d

✖️a

✖️c

✖️b

**9. Mark** all of the following statements that are correct.

✅L2 regularization is a good choice if you don't want any of your parameters to be forced to be

✅L1 regularization is a good choice if you want sparse parameters

✖️Regularization is never used in neural networks

✅Elastic regularization combines L1 and L2 regularization with a mixing parameter

**10.** After the popularization of deep learning, regularization has taken on a broader meaning. Today, regularization does not have to be a specific penalty term added to the loss function, but refers to *any modification we make to the learning that is intended to reduce the test error* (without regard to the training error).

**True** or **false**?

✖️False

✅True

**11.** Consider the following statement about test data.

After training, we can apply our hypothesis h^ to a set of unseen data, the test set, denoted Dtest. The test set is disjoint from the training dataset. We can define the test error for some arbitrary loss function ℓ as 

**Mark** the correct items below.

✖️the sum should be over 

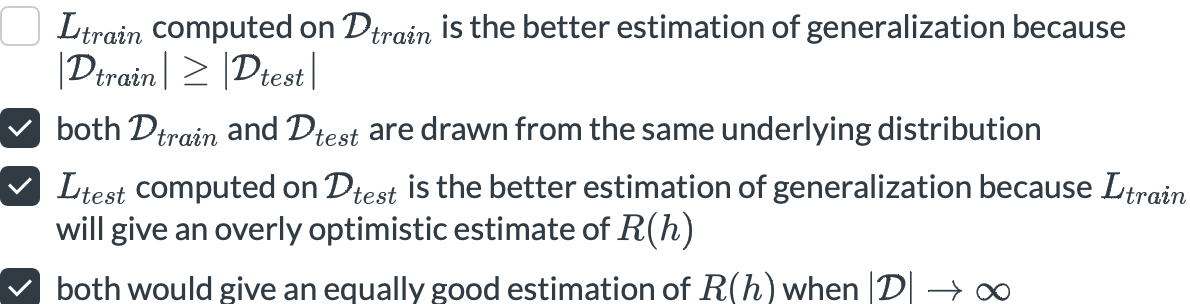
✅the regularization term should not be included in the test error

✅the test error should be normalized by 

✅a disjoint training and test set is written as 

**12.** According to the i.i.d. assumption, for some dataset sampled from P. Both Dtrain and Dtest are sampled from P.

**Mark** the following statement that are correct.



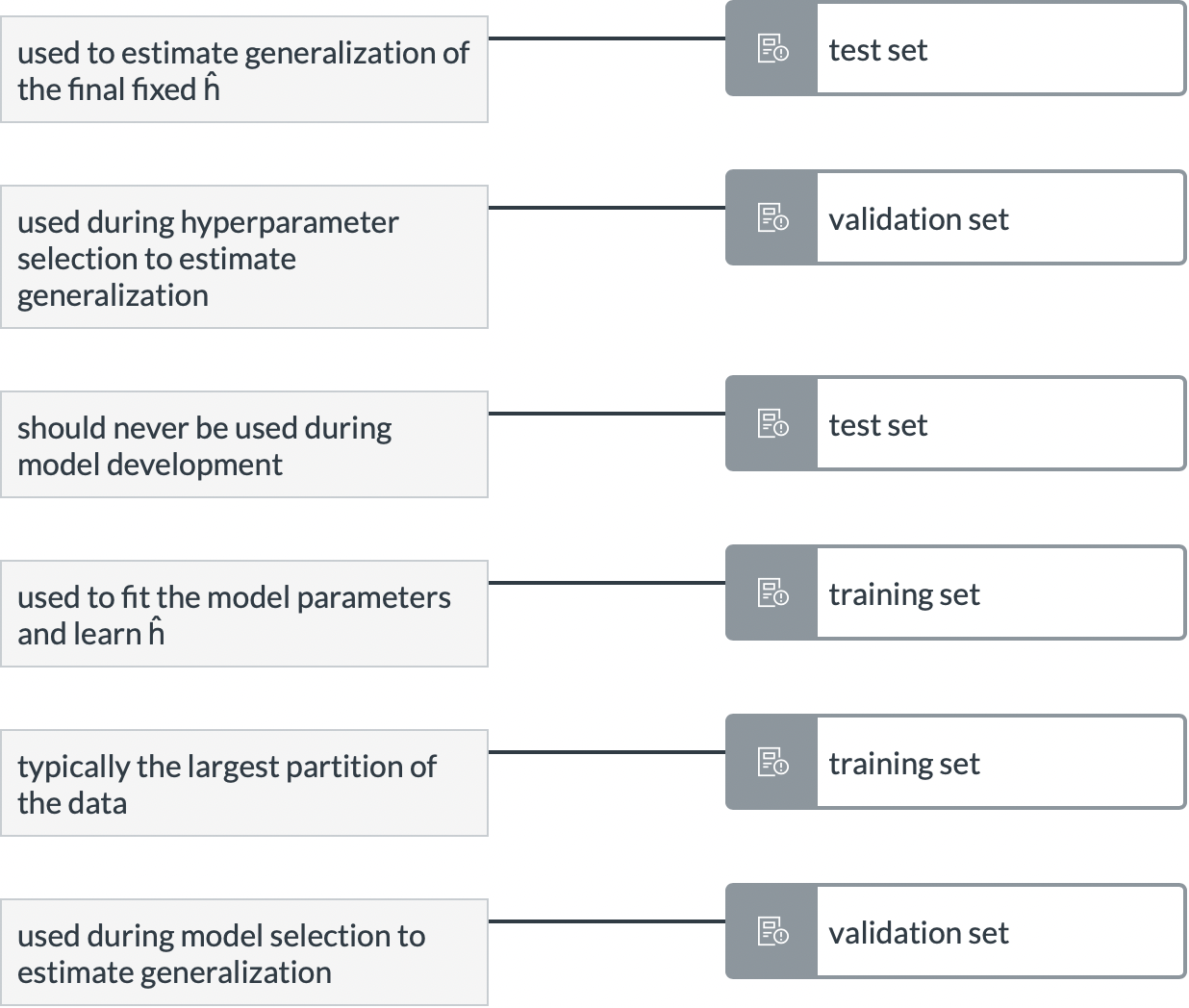
**13.** The **golden rule** for this module on generalization is that you should keep the test set separated and only use it to estimate generalization at the very end, after you have fixed your model.

**True** or **false**?

✖️False

✅True

**14. Match** the following correctly.



**15**. **Fill in** the following statement correctly.

**Model selection** is concerned with choosing the **hypothesis class**(or model type) you want to apply. **Hyperparameter tuning** is concerned with selecting the **parameters** of the model that must be hand-selected. Choosing a feed-forward neural network is an example of **model selection**. Choosing how many **layers** and neurons is an example of **hyperparameter selection**. A **validation set** is necessary for both.

**16.** What are valid reasons for doing cross validation.

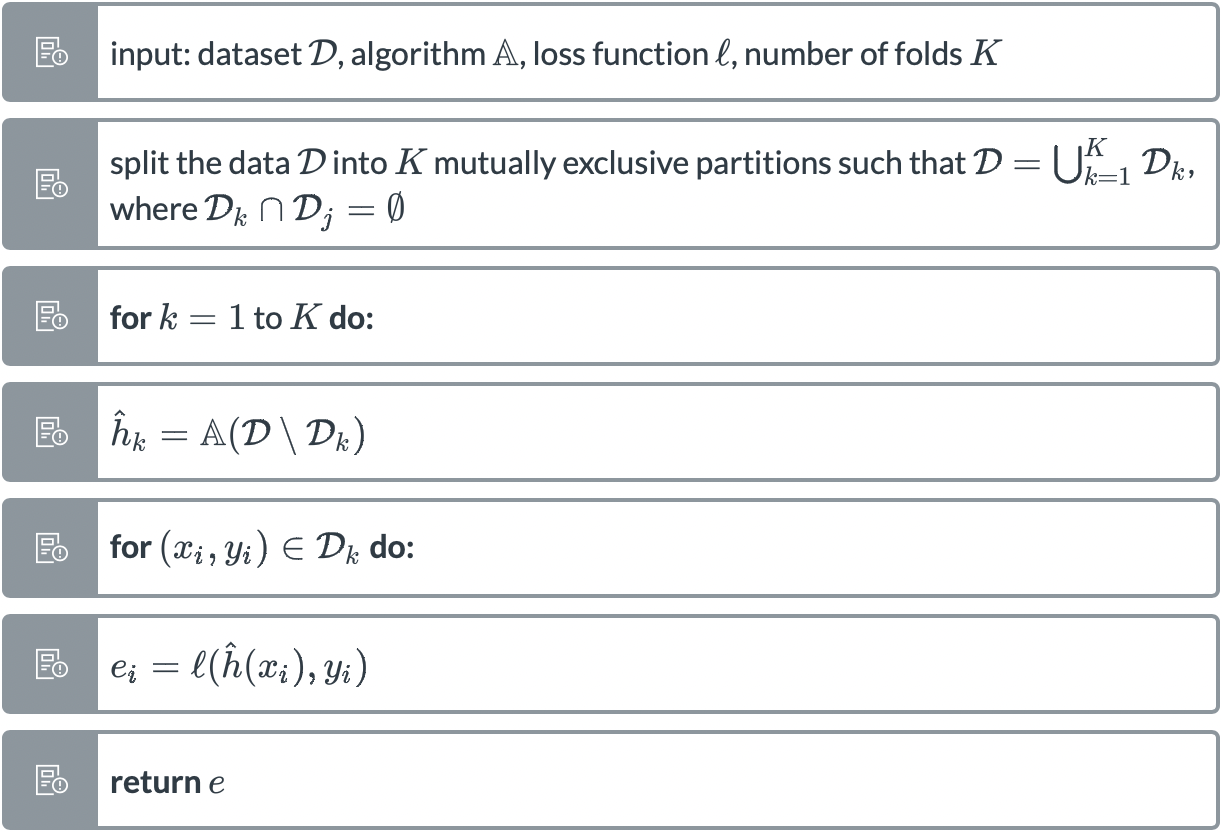
**Mark** all that are correct.

✖️to avoid using a validation set

✖️to reduce computational cost of training

✅If the amount of data is small and we want to have as much to train on as possible

✅To have multiple test error estimates of generalization to compute statistics

**17. Arrange** the following in the correct order for K-folds cross validation.

**18.** The fundamental theorem of PAC learning implies that complex/expressive models should have a worse worst-case upper bound on the 'omniscient generalization gap'. How is it that some methods, *e.g.* deep neural networks with their extremely complex and expressive models, are working so well?

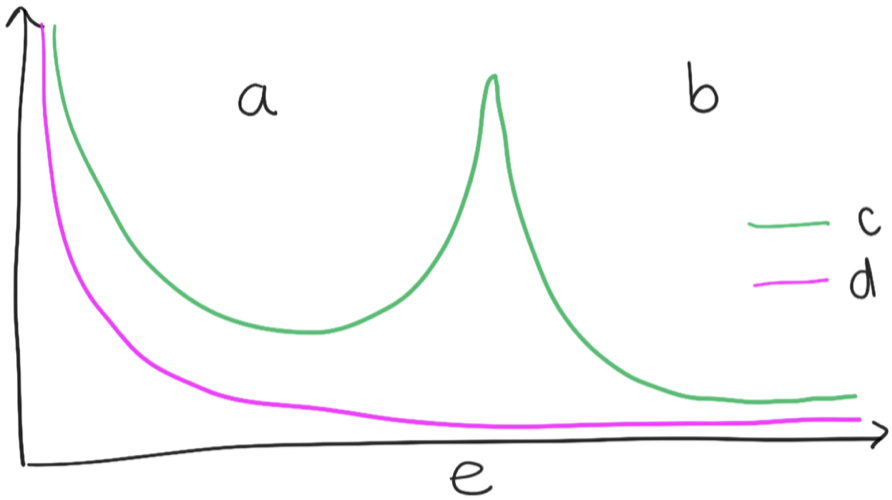
✖️Deep neural networks are just overfitting.

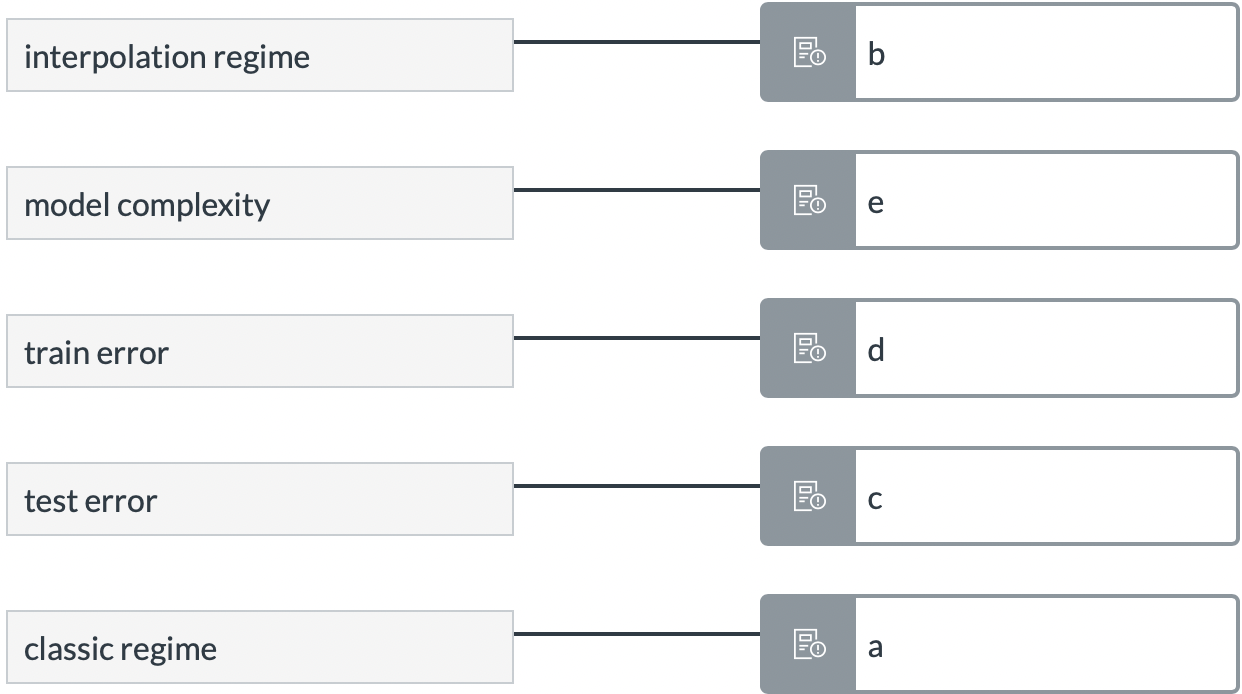
✖️The fundamental theorem of PAC learning does not apply to deep neural networks because they don't have a VC dimension.

✖️The fundamental theorem of PAC learning was disproven in 2012.

✅The fundamental theorem of PAC learning only concerns itself with the worst-case scenario.

**19**. Consider the following figure.



**Match** the following correctly. 

**20.** Which of the following are examples of **data leakage**?

**Mark** all that apply.

✅Normalizing all the data together before splitting it to train, validation, and test

✅Duplicated data appearing in the train/val, val/test, or train/test sets

✅splitting time-series data randomly instead of putting earlier data in train/val and later in test

✅Using the test set to tune model hyperparameters

**21. True** or **false**? Domain shift (or distributional shift) refers to a change between the data distribution the model was trained on and the one it sees when deployed. Domain shifts can cause ML models to behave unexpectedly.

✖️False

✅True

**22. Select** the following situations that are examples of **domain shift**.

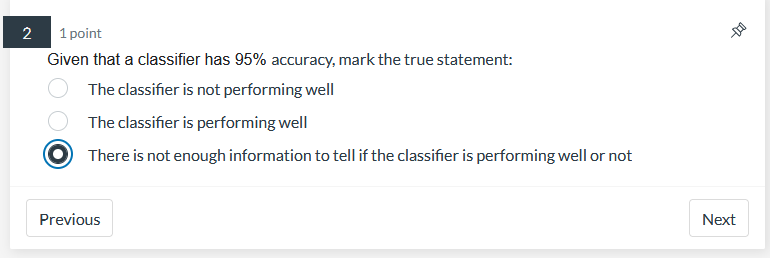
✅A collection of football match results from an entire season up to the championship divided randomly into training and test sets.

✅A dataset of housing prices in Stockholm (training set) and Eskiltuna (test set)

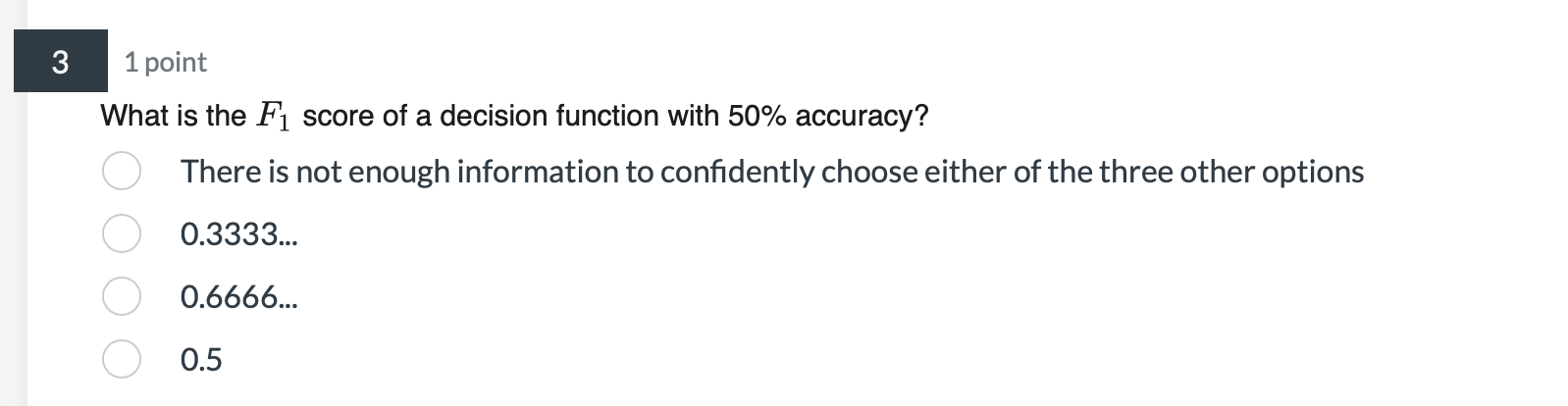
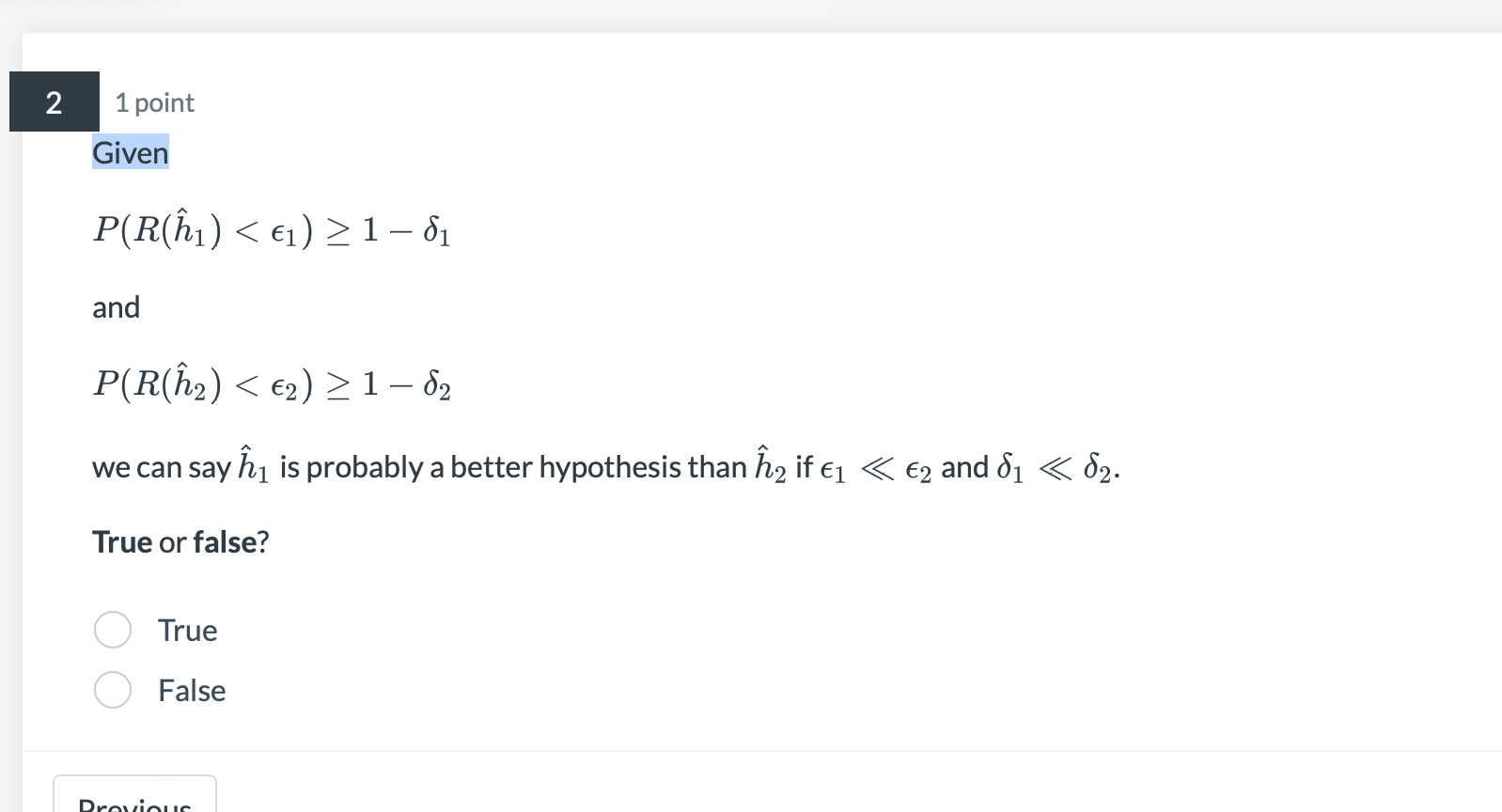
✅A collection of images of street signs from Switzerland (training set) and Sweden (test set)

✅A collection of images of clothing from an online shop (training set) and on people in outdoor environments (test set)

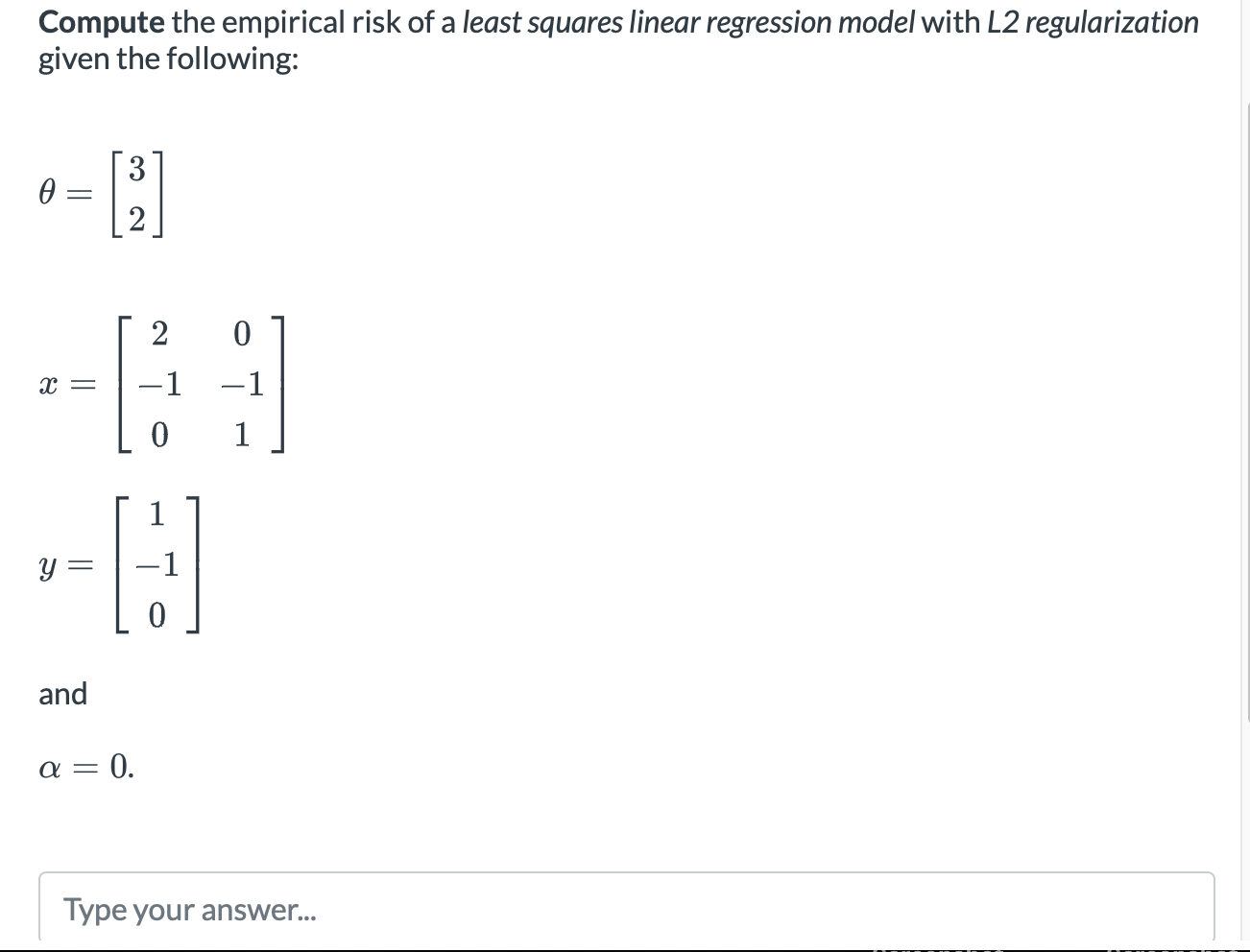
—--------------------------



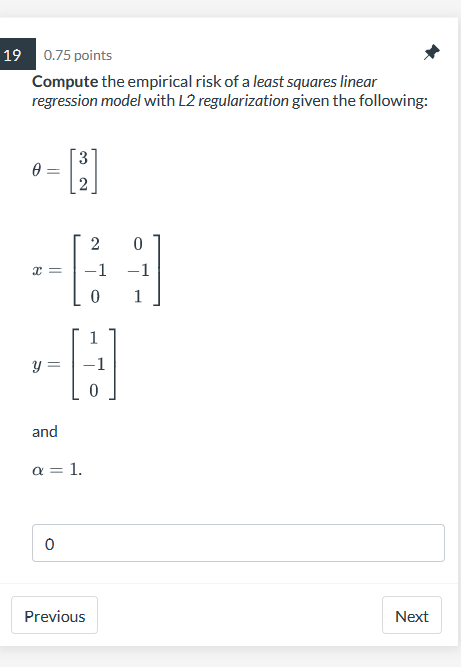
没有

0.5

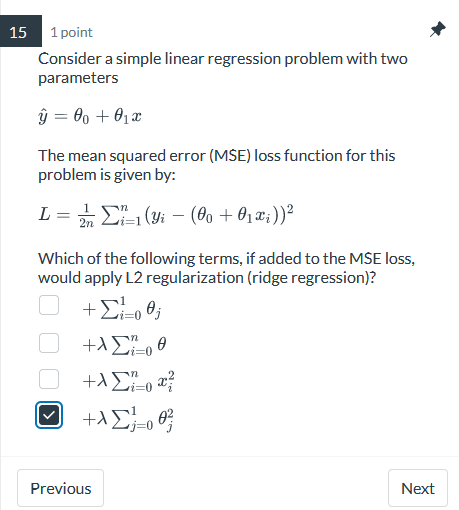
false

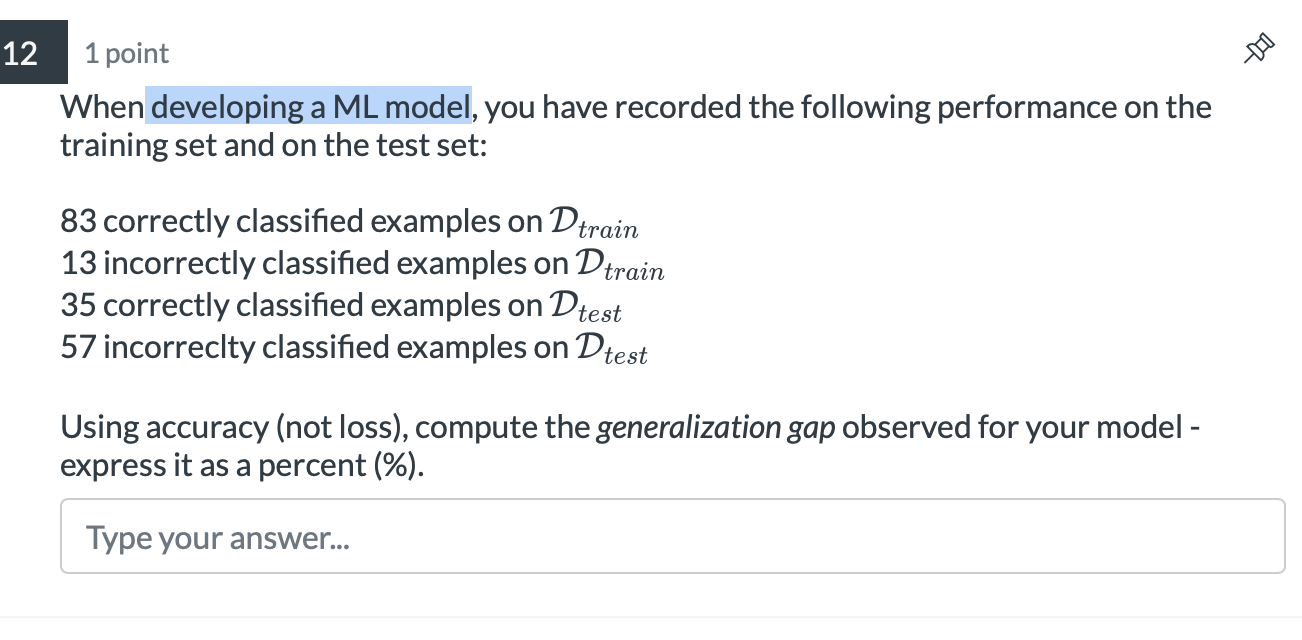


149

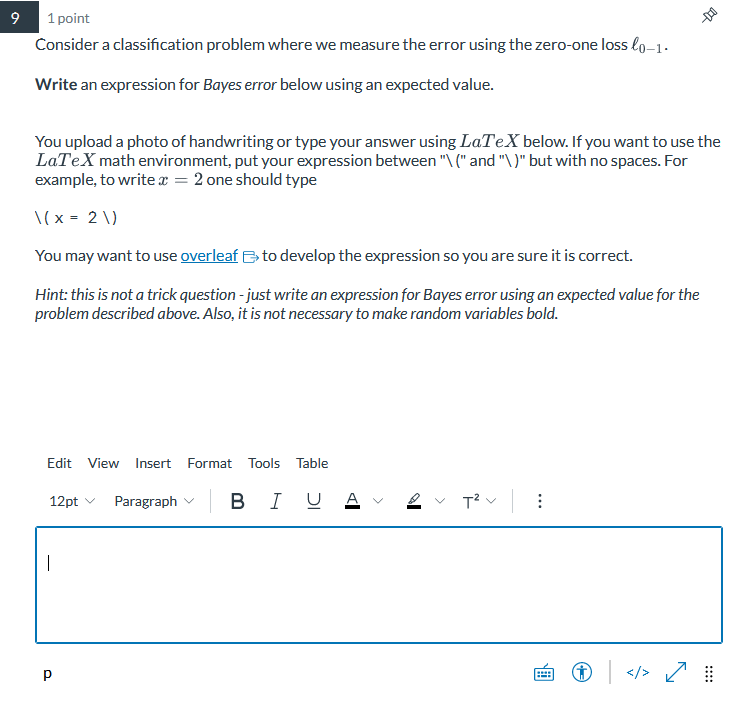


12,67

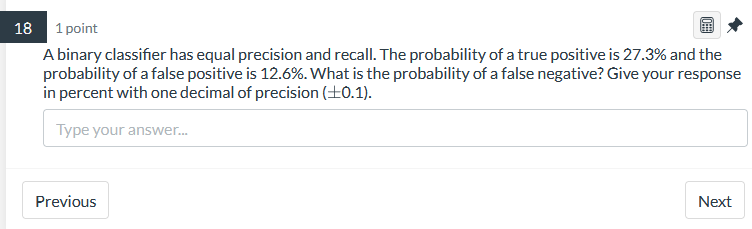


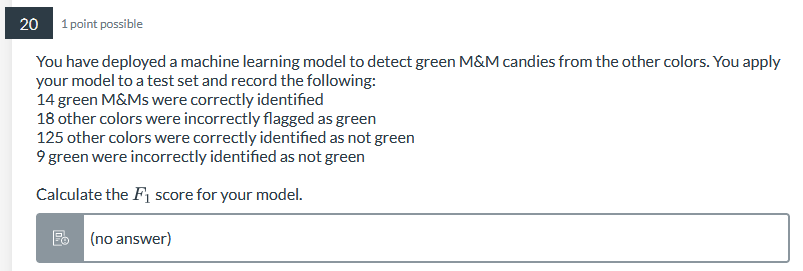


50%%%%%



\[P(\text{error|X}) = \mathbb{E}[l\_{0-1}(Y, h(X))]\]



‘