Midterm Solutions Review

Question 1

- You should answer True or False.
- For False, you have to provide a short correction.

1.1. The difference between supervised learning and unsupervised learning is that the input data in supervised learning is ~~un~~labelled, whereas the input data in unsupervised learning is unlabelled.

99%

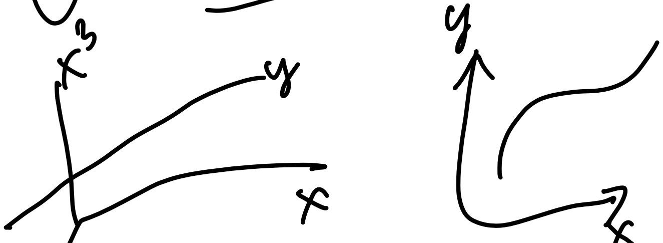
1.2. Polynomial Regression between a scalar input *x* and output *y* is a form of Linear Regression because we establish a linear combination of different polynomial degrees of x.

9290

original feature

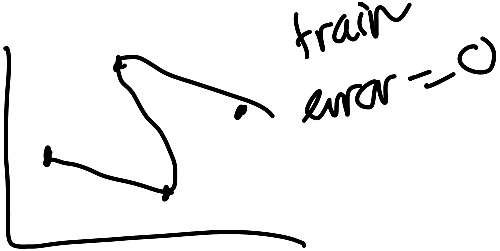
$$y = f(\boxed{x}) = w_0 + w_1 x + w_2 \boxed{x^2} + w_3 x^3$$

non-linear feature



1.3. We should not use the train set when selecting hyperparameters because it is easy to overfit the train set, thus your model might not be able to generalize well on unseen data

92%



1.4. Gradient Descent is an optimization algorithm that can be used to learn the ~~features~~ in Linear Regression and Logistic Regression. weights

~~hyperparameters~~ → not

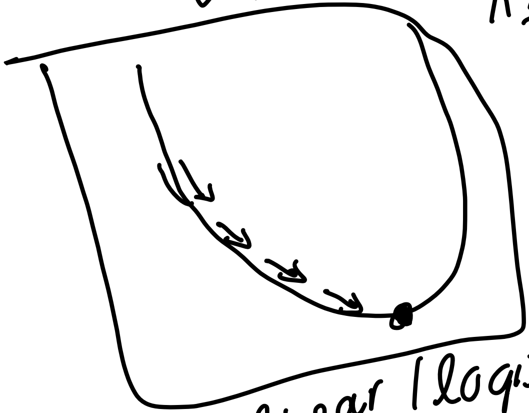
weights | coeffs

features

$$L(w) = w_0 x_0 + w_1 x_1$$

1.5. In gradient descent, at each iteration, we take each step towards the ~~same~~ opposite direction of the gradient (the derivative of the loss function with respect to the weights).

$w -= \lambda \nabla L(w)$



1 local minimum = 1 global minimum

65%

objective for k-means

local

linear / logistic regression

1.6. For the task of predicting house prices on the market, we can use a linear regression model because the output, price of a house, is a discrete value.

increment of 1cm

~~x~~ 69%

regression: predicts continuous value

, ∇

$L(w) =$

1

2

$f(w) = w_0 x_0 + \dots w_D x_D$

1.7. A model with low bias and high variance will generally have a low training error.

999.546 cents

99%

1.8. Before splitting an original dataset into a train set a test set, it is important to randomize it to ensure the samples from both sets are representative of it.

จริง

1.9. An advantage of using the validation set approach compared to the cross-validation one is that the cross-validation set approach is more computationally expensive.

9990

1.10.

1.11. Regularization is a technique that prevents underfitting by decreasing the ~~magnitude~~s of a model's weights to prevent them from being too big.

$|w|$

93%

.negative weigts that are very small

1.12. For a regularized model, as you decrease the regularization parameter, you decrease its bias, thus its training error will be decreased.

92%

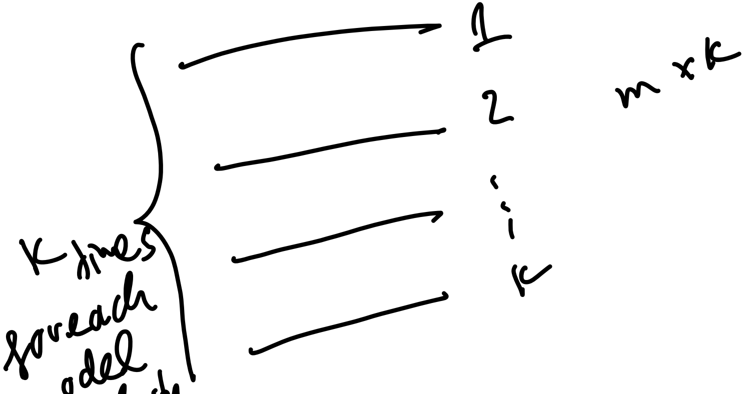
1.13. A greedy algorithm is a problem-solving heuristic that makes a ~~globally~~ optimal choice at each stage.

locally

1.14. One aspect that makes L1 regularization suitable for feature selection is that it can make the weights of a model ~~very close to zero~~.

exactly zero

1.15. k-fold cross validation is more computationally expensive than the validation set approach because if you have *m* model complexities, you only have to train *m* models using the validation set approach, but m * k~~^2~~ models using the k-fold cross validation one.



complexity

1.16. A sigmoid function of an input *x*, can take an input from negative infinity to infinity and outputs a value in the range [-1, 1].

1.17. It's not necessary to normalize numeric features in a decision tree model, as it has no impact on the overall classification error and only the decision boundaries for these features will get scaled differently.

1.18. For a classification task with two output classes (positive and negative), we have $P(+|x) = 1-P(-|x)$ where x indicates an input example.

1.19. When multiplying many probabilities together, we will get a result so small close to 0 that computers might not have the capability of reflecting such a high precision.

1.20. For a classification task with two output classes (positive and negative), if the score of an input xx, defined by Score(x) = w^T x where *w* is a vector of individual weights), is exactly 0, we always predict it to belong to the negative class.

arbitrarily

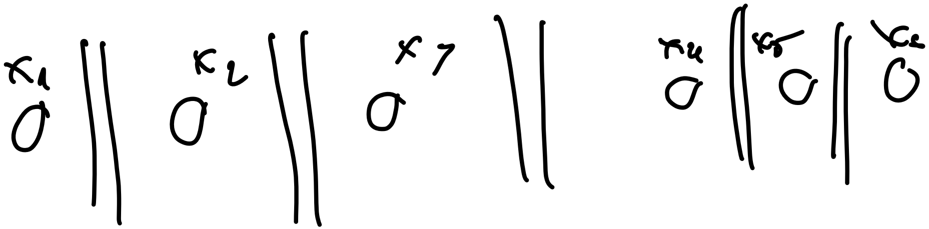
$$\text{Score}(x) \sim [-\infty, \infty]$$

1.21. The true positive rate of a binary classification model in a dataset is defined by the number of true positive examples over the total number of all examples in the dataset.

1.22. It is recommended to have an average scale 1/*n* in defining loss functions (*n* is the number of examples in a dataset), because different datasets have different sizes, so averaging the losses on datasets will make them comparable with each other.

1.23. In Adaboost, the correctly classified examples are given larger weights so that in later iterations they will be paid attention to less.

1.24. In a decision tree, for numeric real-valued features, to choose the best split with the smallest classification error, we can sort all values in descending order and calculate the classification error of the split between every two adjacent values.



1.25. Because Random Forest relies on sampling without replacement to create a separate train set for each decision tree, each of these train sets can have duplicate samples.

1.26. Let $L(w)$ be a loss function on a weight vector w and $R(w)$ be a regularizer on it.
We should use $L(w)$ instead of $L(w) + R(w)$ when calculating the validation and test losses, because $R(w)$ is a penalty applied to control the magnitudes of the weights during training, not a correct measurement of the actual loss.

1.27. We should use Laplace smoothing in Naive Bayes implementation when evaluating on the test set to ensure the conditional probabilities for unseen words (words not appearing in the train set) will not be 0.

1.28. One-hot encoding is for numeric data types.

categorical

nominal vs ordinal

Terrible

ok

good

1.29. In Random Forest, the train set used by each decision tree might contain duplicate samples.

1.30. Adaboost is an ensemble model made up of multiple decision trees, each of which can have many levels.

0

Gaussian

1.31. Adaboost is a generative model.

discriminative

7890

k-means

Question 2: Choose the correct answer.

2.1. Which of the following statements is **correct** about an underfit model?

- High bias, low variance
- Low bias, high variance
- Low bias, low variance
- High bias, high variance

2.2. Which of the following statements is correct for L2 regularization?

- Can be applied to an original loss function of a model using the form $\widehat{w} = argmin_w \, L(w) + \sum_{i=1}^{d} |w_i|^2$
- Tends to have zeros in the solution.
- Is more computationally efficient than L1 regularization because it takes less time to compute dot products.
- Has spikes when visualized in multidimensional space.

2.3. Should we use the train set to fine-tune hyperparameters for model selection?

- ( ) Yes, because the lower the training error is, the better a model can perform.
- ( ) Yes, because the train set is randomly sampled from the original dataset, so train error can be a decent approximation for the true error.
- (X) No, because it is easy to get an extremely low training error with a complex model, but that model might not generalize well on future data.
- ( ) No, because you are supposed to touch the train set once.

2.4. In what scenario is it the most advantageous to use cross-validation over the validation set approach?

- When there are not a lot of samples in the training data.
- When there are a significantly large number of samples in the training data
- When there are limited computing resources.
- When the models are currently underfitting.

2.5. A model is trained on a dataset. We notice that the magnitudes of individual weights are very big compared to others. Without additional information, which of the following is the correct assessment of this issue?

you've not normalized the dataset yet

- This is always a sign of overfitting. ←
- If the dataset is not normalized, the large weights might result from features with naturally small units.
- We should normalize the original dataset before splitting it into train set and test set.
- We do not have to normalize the dataset, as regularization can scale the features appropriately.

2.6. Which of the following statements is not correct regarding classification?

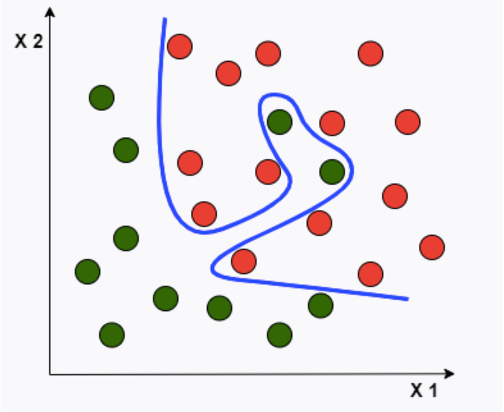
- It is not possible to use gradient descent with the MSE loss function on the classification task because the loss function in this case isn't differentiable.
- A good baseline model for the classification task that involves many output classes is to always predict the output class that has the least training examples.
- In logistic regression, we maximize the joint conditional likelihood of assigning all training inputs to their correct classes.
- For a binary classification task, false negatives and false positives might have different degrees of practical significance depending on the circumstances.

2.7. Which of the following is not correct regarding performance assessment of a model?

- The errors of a model consist of bias, variance and irreducible errors.
- The bias-variance tradeoff is formulated in regards to the true error, and there exists an optimal choice for a model complexity with a minimized true error where both bias and variance are not too high or low.
- A model that pays too much attention to small fluctuations and outliers in the training set shows a strong sign of overfitting.
- It is impossible to achieve an exact training error of 0, because there exists an irreducible error in every model.

2.8. Which of the following is the most correct description for the following image, which shows the decision boundary for a classification model on a training set with two features x1 and x2?

- It doesn't represent the decision boundary for a decision tree.
- The model is likely to have a high accuracy on the test set.
- This illustrates an example of underfitting.
- If the model is currently regularized, to improve the predictive performance of this model, we should decrease the regularization parameter.



2.9. Which of the following is not correct about the log trick that we have learned so far?

- It is commonly used in Decision Trees.
- When calculating the MLE, we use it to turn a product of probabilities into a sum of log probabilities.
- It makes computing the derivative of a product with multiple terms easier.
- It helps avoid floating point overflow / underflow issues in computers.

2.10. Which of the following is correct about decision tree?

- Categorical features always have to be encoded before being used in the model.
- Its decision boundaries are often difficult to understand. interpretable
- [✓ The algorithm that it uses to expand the tree doesn't guarantee a globally optimal decision at each step.] greedy + recursive
- At each leaf node, it can recursively build the next stump by choosing a split on a feature that yields the smallest classification accuracy.

2.11. Which of the following is **correct** about Adaboost?

- ( ) It can be trained in parallel.
- ( ) It will never overfit.
- ( ) Each of its individual model makes an independent and equal decision in the final prediction.
- (X) Each of its individual model, a stump, is a weak learner as each stump has high bias.

2.12. Given a particular model, which of the following most likely **will not** resolve the overfitting issue?

- (X) Decreasing the number of training examples
- ( ) Using a more simple model
- ( ) Decreasing the number of features
- ( ) Applying some regularization to the model

2.13. Which of the following, if applied, will not cause any change to the performance of an unregularized linear regression model?

- ( ) Increasing the learning rate in gradient descent
- (X) Changing the unit of a feature
- ( ) Using a different loss function (i.e. Mean Square Error -> Mean Absolute Error)
- ( ) Using a different train set

a8%

$$L(w) = w_0 \underline{\overset{m \to km}{x_0}} + w_1 x_1 + (w_0^2 + w_1^2) = 0$$

$$= w' + w'_1 x_1 + (1000^2 w_0^2 + w_1'^2) = 0$$

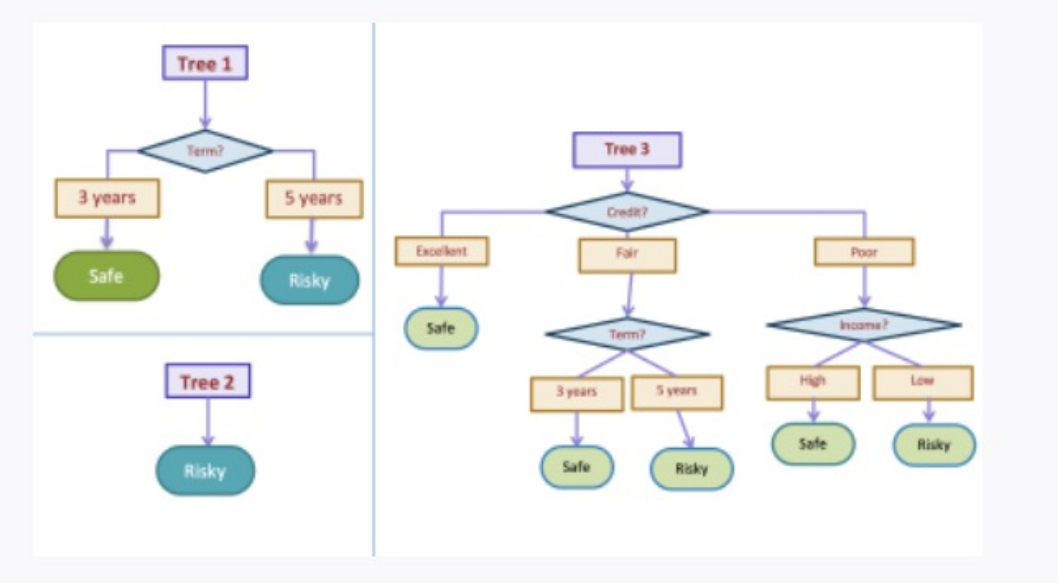
$$L(w) =$$

$$w'_0 = 1000 w_0$$

2.14. Which of the following classification model does not involve creating decision boundaries in the data space?

- (X) Naive Bayes
- ( ) Logistic Regression
- ( ) Adaboost
- ( ) Random Forest

2.16. Which individual decision tree is **most likely** to have the highest bias?



Question 3: Select all that apply.

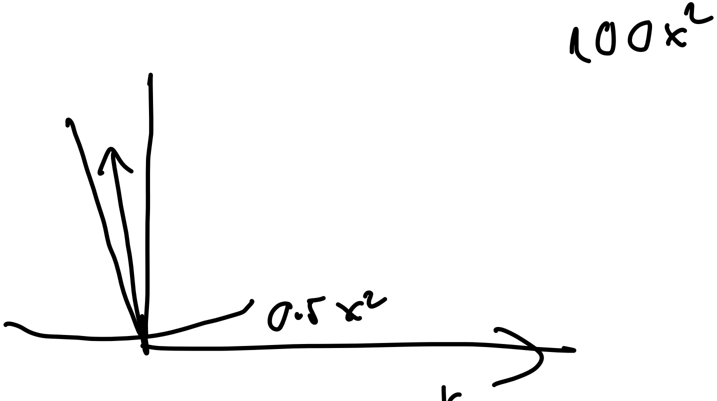
3.1. What of the following does not describe a classification task?

- [ ] Detecting a person's facial expression
- [ ] Diagnosing whether someone has a cancer or not
- [X] Clustering news articles into different categories without any labelled tags
- [X] Predicting the cost of living of cities around the world

3.2. As you increase the number of features, what of a model will also generally increase?

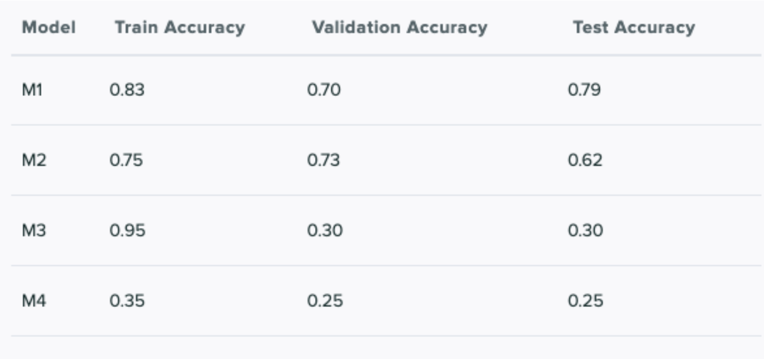
- [X] Complexity
- [ ] Bias
- [X] Variance
- [ ] ~~Weights~~

Magnitude of the largest weights



3.3. Four different classifications models are trained and evaluated but with different sets of features; each one is labeled M1 through M4. Assume the training set, validation set, and test set are the same for each model. What are the correct assessments of the models?

- [ ] M4 probably underfits because it has a very low ~~validation~~ training classification error.
- [X] M3 probably has a high variance because it has low classification error on the train set but high error on the validation one.
- [X] M2 probably has a lower bias than M4.
- [ ] We should choose M1 out of the 4 models to perform on unseen datasets because it has the highest ~~test~~ validation accuracy.



Model	Train Accuracy	Validation Accuracy	Test Accuracy
M1	0.83	0.70	0.79
M2	0.75	0.73	0.62
M3	0.95	0.30	0.30
M4	0.35	0.25	0.25

3.4. What are correct about regularizing a model?

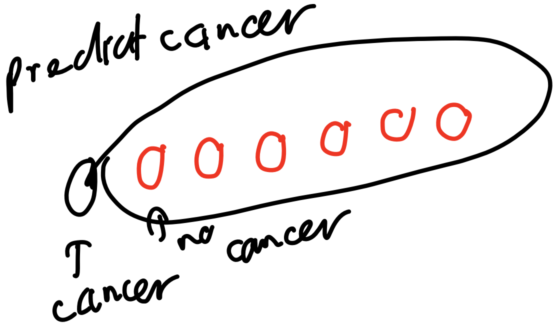
- [X] We should normalize the features to make sure they are on the same scale.
- [ ] Decreasing the regularization term prevents a model from overfitting.
- [X] We should not regularize the bias (intercept) term because it is a linear scale and doesn't affect a model's complexity.
- [ ] $R(w)$ = $|\sum_{j=1}^{d} w_j|$(D is the number of a features of a model) is a possibly good choice as a regularizer for the loss function $\widehat{w}$ = $argmin_w\, L(w) + R(w)$

3.5. What are the correct assessments of various classification models?

- [X] Decision Tree and Logistic Regression are both discriminative classifiers.
- [X] It is mandatory to encode categorical features into numeric types before training a Logistic Regression model.
- [X] One important assumption in Naive Bayes is that it treats each word in a sentence independently of the others, but in practice words derive meaning from surrounding contexts.
- [X] An advantage of using Decision Tree over other classification models is due to how interpretable it is.

3.6. What are properties of both Linear Regression and Logistic Regression models?

- [X] They are both supervised learning tasks.
- [X] Applying regularization (L1, L2) are possible way to prevent them from overfitting.
- [ ] We can optimize them using Gradient Descent with MSE loss function.
- [X] When used on the space of the transformed feature values, they can be represented by a linear line (Linear Regression is a regression line and Logistic Regression has a linear decision boundary).



3.7. What are not correct about a Logistic Regression model?

- [ ] It uses the sigmoid function to turn the score of an input into a probability form.
- [X] It has a closed-form solution.
- [ ] Its loss function is continuous and differentiable.
- [ ] It uses maximum likelihood estimation (MLE) definition in calculating the loss.

Logistic regression

NLL error (continuous)

Classification error = $\frac{\text{Incorrect Predictions}}{\text{All examples}}$

Discrete

error = 100

3.8. What are correct about classification error / accuracy

- [ ] It is possible for a Linear Regression model achieve a training accuracy of 100%.
- [ ] The negative log-likehood loss function in Logistic Regression aims to minimize the total classification error on the training set.
- [X] Depending on contexts, a test classification error of 0.001% might not always depict the full picture of how good a model is.
- [X] The classification accuracy and error rate are complements of each other, meaning that 1 - accuracy = error

everyday objects

pencil: $5 predict $1

house 5 million predict: 4.8 million

price of a house is 5k 400k

1 million

3.9. What are correct about a decision tree?

- [X] A decision tree has a higher chance of overfitting than a random forest.
- [ ] The more levels a tree has, the higher the bias is.
- [ ] It is a generative model.
- [X] It uses a greedy and recursive algorithm.

3.10. Between Adaboost and Random Forest, select all properties that are unique to only Adaboost:

- [ ] It is an ensemble model composed of many individual models.
- [ ] The training process can be sequential (training individual models according to a certain order, usually using a single machine).
- [X] From the initial model, the more models being added to the ensemble, the higher chance of overfitting there is.
- [ ] The prediction of each individual model is independent of the others.

for tree in self._trees:

3.11. What of the following assessments are **correct** regarding some hyperparameters:

- [X] There is lot of unknown when choosing hyperparameters because they can’t be directly learned during training.
- [X] We can use the validation set approach or cross-validation to fine-tune hyperparameters.
- [ ] In gradient descent, the larger a learning rate is, the better, because it significantly reduces the convergence time for the model.
- [X] For a decision tree, when we decrease the minimum number of examples to split in a leaf node, the tree can grow larger and lead to a higher chance of overfitting.