DS-GA-1003 Spring 2020 Final Exam

Name and section: ______________________________

1 True or False

1. (1 point) Adaboost usually works better than random forests if there are outliers in the dataset.

   ○ True ✓ **False**

2. (1 point) [SKIP] Bagging reduces variance of decision trees.

   ✓ **True** ○ False

3. (1 point) Logistic regression can be kernelized.

   ✓ **True** ○ False

4. (1 point) Perceptron will find a unique solution if the data is linearly separable.

   ○ True ✓ **False**

5. (1 point) SGD will always converge to a local optimum when learning neural networks.

   ○ True ✓ **False**

6. (1 point) Given an SVM with quadratic kernel, the dual form is more efficient at inference time than the primal form, assuming the number of support vectors is smaller than the input feature dimension.

   ✓ **True** ○ False

   Solution: Dual form taks $O(np)$ and primal form takes $O(p^2)$ where p is the input dimension and n is the number of support vectors.

7. (1 point) The concepts of bias and variance exist in both Bayesian and frequentist approaches.

   ○ True ✓ **False**

8. (1 point) Random forest is more efficient than gradient boosted trees because it can be parallelized during training.

   ✓ **True** ○ False

9. (1 point) Using more features helps the learning algorithm generalize better to unseen data.

   ○ True √ **False**

10. (1 point) $L1$ regularization can be used for feature selection.

    √ **True** ○ False

11. (1 point) Multiclass SVMs use the same weight vector for all classes.

    √ **True** ○ False

    Solution: Multiclass SVM uses generalized hinge loss that considers a single weight vector.

12. (1 point) In bagging, increasing bootstrap sample size can reduce variance.

    ○ True √ **False**

13. (1 point) Absolute loss is more robust to outliers than squared loss for regression.

    √ **True** ○ False

14. (1 point) Backpropagation uses dynamic programming.

    √ **True** ○ False

15. (1 point) Given a dataset $\{x_1, x_2, \ldots, x_n\}$ sampled from a data generating distribution P, x_1 is an unbiased estimate of the mean of the distribution.

    √ **True** ○ False

2 Multiple Choices

Note: There can be more than one correct answers; select all that apply! No partial credits: all correct answers must be checked.

1. (3 points) Which of the following models can be learned by MLE?

   ○ Perceptron ○ Ridge regression ○ SVMs √ **Logistic regression**

   Solution: Ridge regression is learned by MAP. The others aren't probabilistic models.

2. (3 points) If we find our learned model is overfitting, what should we do?

   ○ Increase model complexity. √ **Increase regularization.**
   √ **Add more data.** ○ Training for a longer period of time.

3. (3 points) Which of the following is true about support vectors?

   ○ Support vectors are the examples closest to the decision boundary.
   ○ Removing support vectors during training will not change the decision boundary.
   √ **Support vectors are the only examples needed for inference.**
   √ **There are at least two support vectors.**

   Solution: Examples within the margin can be closer to the decision boundary than the support vectors.

4. (3 points) In a classification setting, which of the following are usually used as weak learners for boosting?

   ○ A classifier that always predict the majority class.
   √ **A decision stump.**
   ○ A classifier that uniformly predicts a class at random.
   ○ Perceptron.

   Solution: A is a weak learner but its output is constant, thus it doesn't make sense to use for boosting. D is a linear function and weighted combinations of a linear function is also linear, so it wouldn't be different from a linear model thus is not usually used.

5. (3 points) Which of the following description is true about kernels?

   ○ A kernel defines a feature map ϕ that maps x to an infinite-dimensional space.
   √ **The kernel function k can be considered as a similarity function for two data points.**
   √ **The kernel trick provides an efficient way to compute inner products in high dimensional space.**
   √ **A valid kernel function must produce a positive semi-definite kernel matrix.**

6. (3 points) Which of the following are appropriate activation functions for neural networks?

   ○ $2x$ √ $\max(x, 0)$ ○ $\begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$ √ $\begin{cases} x & x > 0 \\ 0.01x & x \leq 0 \end{cases}$

   Solution: Nonlinear and differentiable.

7. (3 points) Which of the following can reduce variance of a learning algorithm?

○ Increase the maximum depth of a decision tree.
○ Increase the variance of a zero-mean Gaussian prior of the parameters.
○ Increase the step size (shrinkage) in gradient boosting.
√ **Increase λ (weight on the penalty term) in ridge regression.**

8. (3 points) Which of the following loss functions is an upper bound of the 0-1 loss function? (Let $y \in \{1, -1\}$ be the gold label and $f(x) \in \mathbb{R}$ be the predictor output.)

√ $(1 - yf(x))^2$ ○ $\max(0, -yf(x))$ √ $e^{-yf(x)}$ ○ $\log(1 + e^{-yf(x)})$

Solution: D can also be correct as the base of the log is not specified here.

9. (3 points) Consider a small XOR dataset:

x_1	x_2	y
+1	+1	0
+1	-1	1
-1	-1	0
-1	+1	1

Which of the following model can achieve zero training error?

○ A decision tree with depth 1.
√ **A decision tree with depth 2.**
√ **A two-layer neural network with 2 hidden units and ReLU activation function.**
○ Linear soft-margin SVM.

10. (3 points) If the solution of linear regression on a dataset is w^*, what solution would you get if you scale all features by a factor of c (i.e. replace x by cx) before doing regression?

○ w^* √ $\frac{w^*}{c}$ ○ $\frac{w^*}{c^2}$ ○ cw^* ○ c^2w^*

11. (3 points) Assuming that we can always obtain the optimal solution, adding a regularization term to the objective function of logistic regression can

○ Decrease training error √ **Increase training error**
√ **Decrease validation error** √ **Increase validation error**

12. (3 points) Which of the following model will always have a unique optimum regardless of the data?

○ Linear regression √ **Ridge regression**
○ Logistic regression ○ Lasso regression

Solution: Other models don't have a unique solution when

- Linear regression: X is not full-rank.
- Logistic regression: data is linearly separable (see hw5).
- Lasso regression: duplicated features (see lec4 slides).

13. (3 points) Which of the following are necessary conditions for the minimum risk of a classification problem using 0-1 loss (i.e. the Bayes error rate) to be zero?

√ **The distributions $P(X \mid Y)$ don't overlap.**
○ The training data is linearly separable.
○ Cannot tell. It depends on the class prior $P(Y)$.
○ $X \mid Y$ follows a Gaussian distribution.

14. (3 points) Which of the following is true about neural networks?

○ The objective function is convex.
√ **Their nonlinearity is due to the nonlinear activation functions.**
○ They can only be used for classification.
√ **Yann Lecun won Turing Award because of his contribution to neural networks.**

15. (3 points) Consider OvA and AvA for multiclass classification, if the base binary classifier takes $O(N^2)$ time to train, where N is the number of training examples, which method is more efficient in terms of training time (assuming no parallelization)?

○ OvA √ **AvA** ○ Equally efficient

Solution: OvA time: $O(kN^2)$. AvA time: $O(k^2(N/k)^2) = O(N^2)$.

16. (3 points) Assuming that the data is not separable, increasing C in soft-margin SVMs can

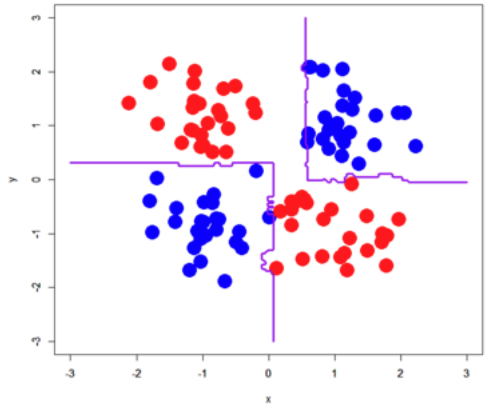
√ **Reduce margin** ○ Increase margin
√ **Reduce training error** ○ Increase training error

17. (3 points) Which of the following is true about the Naive Bayes classifier?

√ **It assumes features are independent conditioned on the label.**
○ It only works with categorical features.
○ It has a linear decision boundary.
√ **It is a generative model.**

Solution: Gaussian naive bayes does not have linear decision boundary in general (see lec09 slide 44).

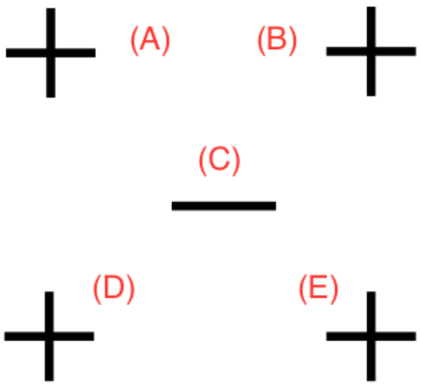
18. (3 points) Which model could have generated the decision boundary shown in the figure below?



○ SVM with Gaussian kernel ○ Logistic regression
○ Neural networks √ **Random forest**

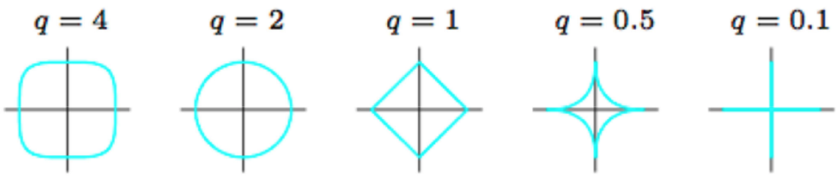
Solution: Decision boundary of neural networks will be smooth, not piecewise constant with the little "bumps".

19. (3 points) Consider training an Adaboost classifier using decision stumps. Given the following 2D dataset, which examples will have their weights increased after the first iteration?



○ A ○ B √ **C** ○ D ○ E

20. (3 points) Consider the regularizer $\ell_q = \left(\sum_{i=1}^{d} |w_i|^q\right)^{\frac{1}{q}}$. For $d = 2$, which of the following will produce sparse weights?



○ $q = 4$ ○ $q = 2$ ✓ $q = 1$ ✓ $q = 0.5$ ✓ $q = 0.1$

3 Short Questions

1. **Optimization.** Consider $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = \max(x^2, 2x)$.
   (a) (1 point) Is $f(x)$ convex?
   ✓ **True** ○ False
   (b) (2 points) At which points does the function have non-unique subgradients?

   Solution: $\{0, 2\}$.

   (c) (3 points) What is the subgradient at $x = 2$? Show all valid subgradients if there is more than one.

   Solution: $[2, 4]$.

   (d) (1 point) How many local minima does $f(x)$ have?

   Solution: One.

2. **MLE and Bayesian methods.** To estimate the number of visitors to a store (denoted by X) during some fixed time of the day, we assume that X follows a Poisson distribution:

$$p(X = k \mid \lambda) = \text{Poisson}(\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (\lambda > 0).$$

   We then recorded the number of visitors for 10 days and got the following data points:

$$\mathcal{D} = (5, 3, 12, 7, 6, 8, 6, 10, 9, 9).$$

   (a) (4 points) What is the maximum likelihood estimate of λ given the observed data?

   Solution: $\lambda_{\text{MLE}} = \sum_{i=1}^{n} x_i / n = 7.5$.

   (b) (4 points) Let's put a gamma prior on λ:

$$p(\lambda) = \text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \quad (\lambda, \alpha, \beta > 0),$$

where $\Gamma(\alpha)$ is the gamma function (you can treat it as a constant). If we take the prior Gamma$(5, 1)$, what is the posterior $p(\lambda \mid \mathcal{D})$?

Solution:

$$p(\lambda \mid \mathcal{D}) = \text{Gamma}(\alpha + \sum_{i=1}^{n} x_i, \beta + n) = \text{Gamma}(5+75, 1+10) = \text{Gamma}(80, 11).$$

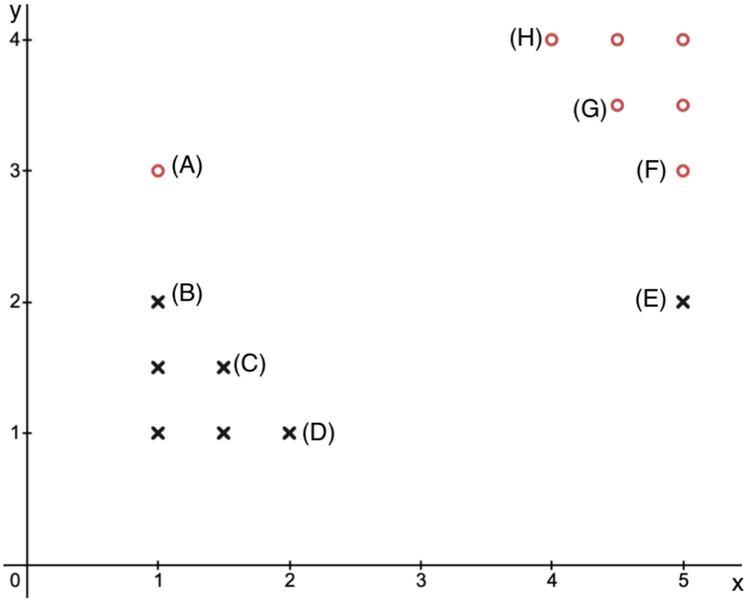
(c) (1 point) Is the Gamma prior conjugate to the Poisson model?

√ **True** ○ False

(d) (2 points) Suppose we need to update our estimate of λ as more data arrives, what is the amount of data storage needed if the total number of data points is n?

Solution: $O(1)$. We just need to save the sum and n.

3. **SVMs.** Consider the 2D datasets shown in the figure below. We want to classify the red circles and the black crosses using SVMs.



(a) (1 point) Is the data linearly separable?

√ **True** ○ False

(b) Let's first consider using a linear hard-margin SVM.

i. (3 points) What's the learned decision boundary?

Solution: $y = 2.5$.

ii. (2 points) Select all points that are support vectors.
✓ **A** ✓ **B** ○ C ○ D ✓ **E** ✓ **F** ○ G ○ H

iii. (2 points) What's a potential problem of hard-margin SVMs (as demonstrated in this example)?

Solution: Not robust to outliers or label noise.

(c) Next, let's consider a linear soft-margin SVM.

i. (3 points) Assuming C is not too extreme (very small or very large), what is a likely decision boundary learned by the soft-margin SVM? (The answer doesn't have to be exact as long as it shows a qualitative difference from the hard-margin SVM.)

Solution: E.g., $y = 5 - x$.

ii. (2 points) What's the training error rate of the learned SVM?

Solution: $2/14 = 1/7$.

(d) (3 points) Suppose we got the optimal solution w_1 of the hard-margin SVM on the dataset. If we add a new feature to each data points and compute the new optimal solution w_2, which of the following might be true?
○ $\|w_2\|_2 > \|w_1\|_2$ ✓ $\|w_2\|_2 = \|w_1\|_2$ ✓ $\|w_2\|_2 < \|w_1\|_2$

Solution: Adding a feature cannot increase the objective because if so the model can put zero weight on that feature.