STAT 451 Exam1

1. If a question is ambiguous, resolve the ambiguity in writing. We will consider grading accordingly. e.g.

   - In #10, I think "average" refers to the population mean μ (not the sample mean $\bar{X}$).
   - In #13b, I think ...

   Please answer this question with a period (.) if you have no other comment, so that Canvas will think you answered it and give you its point(s). Do not write unnecessary comments.

Answer each question at https://canvas.wisc.edu/courses/408488/quizzes/541565 **as you work through the exam** so that you do not run out of time with questions unanswered.

2. Consider 3-NN (three nearest neighbors) using the Minkowski distance with $p = 1$.

   (a) Find the distance from $\mathbf{z} = (3, 4)$ to each of the other points $\mathbf{x}$:

$\mathbf{x}$	y	Distance from $\mathbf{z}$ to $\mathbf{x}$
$(-1, -1)$	1	
$(0,\ 1)$	0	
$(1,\ 0)$	0	
$(2, 3)$	1	

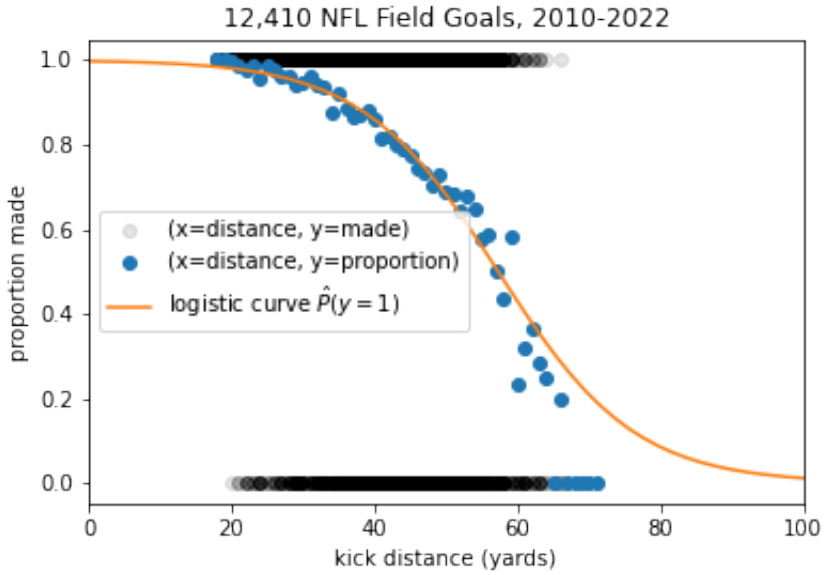
   (b) How does 3-NN classifiy $\mathbf{z}$?

   (c) What y value does 3-NN regression predict for $\mathbf{z}$?

   (d) How does weighted 3-NN classifiy $\mathbf{z}$?

3. Consider a logistic regression model with $\mathbf{w} = (1, 2)$ and $b = 0$.

(a) From the logistic regression model represented in the figure, estimate the likelihood of an NFL field goal kicker making three field goals in a row, one from 20 yards, one from 40, and one from 60. We may suppose these attempts are independent and make other reasonable simplifiying assumptions.



P(NFL kicker makes the three field goals) $\approx$ _____. (Give a reasonable estimate.)

(b) In calculating the coefficients for a logistic regression model, why do we minimize negative log likelihood instead of maximizing likelihood? Mark each statement as a true or false reason by circling the appropriate choice.

i. TRUE / FALSE A product of probabilities can overflow in fixed-precision computer arithmetic.

ii. TRUE / FALSE A product of probabilities can underflow in fixed-precision computer arithmetic.

iii. TRUE / FALSE The natural log of a product is naturally expressed as a sum, and differentiating a sum is easier than differentiating a product.

iv. TRUE / FALSE The natural log is strictly increasing, so maximizing the likelihood is the same as minimizing the negative log likelihood.

v. TRUE / FALSE Using the natural log facilitates finding a closed-form expression for the coefficients in terms of the data.

4. Here are some questions on support vector machines.

   (a) Suppose we have a soft-margin SVM for which $\mathbf{w} = (2, 3)$ and $b = -1$. How does the SVM classify $(\mathbf{x} = (2, 1), y = 1)$?

   (b) Suppose we have some training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ in matrices `X` and `y`. We have plotted the data with $y = -1$ examples red and $y = 1$ examples blue. Which line of code gives the best model for predicting new examples?

   For each question, write the best answer from among these lines labeled "A" thorugh "H". You may not use an answer more than once.

   A: `clf = svm.SVC(kernel='linear', C=1); clf.fit(X, y)`

   B: `clf = svm.SVC(kernel='linear', C=1000); clf.fit(X, y)`

   C: `clf = svm.SVC(kernel="rbf", C=1, gamma=1); clf.fit(X, y)`

   D: `clf = svm.SVC(kernel="rbf", C=1, gamma=10); clf.fit(X, y)`

   E: `clf = svm.SVC(kernel="euclidean", C=1, gamma=2); clf.fit(X, y)`

   F: `clf = svm.SVC(kernel="euclidean", C=1000, gamma=2); clf.fit(X, y)`

   G: `clf = svm.SVC(kernel="gini", C=1); clf.fit(X, y)`

   H: `clf = svm.SVC(kernel="gini", C=1000); clf.fit(X, y)`

   i. _____ The data are two linearly-separable clouds of points, one red and one blue.

   ii. _____ The red points are scattered between $x = -3$ and $x = 3$ and roughly along $y = x^2$, a parabola with vertex $(0, 0)$ that opens up. The blue points are scattered between $x = -3$ and $x = 3$ and roughly along $y = x^2 + 2$, a parabola with vertex $(0, 2)$ that opens up.

   iii. _____ The data consist of two clouds of points, one red and one blue, that are linearly-separable except for a few outliers of each color.

   iv. _____ The data consist of mixed red and blue points scattered randomly in the region $x \in [0, 1]$ and $y \in [0, 1]$.

5. Mark each statement TRUE or FALSE by circling the appropriate choice.

   (a) TRUE / FALSE In linear regression, a reasonable alternative to the cost function *mean squared error* $= \frac{1}{N}\sum_{i=1}^{N}[f_{\mathbf{w},b}(\mathbf{x}_i) - y_i]^2$ is *sum of squared error* $= \sum_{i=1}^{N}[f_{\mathbf{w},b}(\mathbf{x}_i) - y_i]^2$.

   (b) TRUE / FALSE Using gradient descent to minimize $z = f(\mathbf{x})$ can be slow because the algorithm requires many calls to f, which is expensive if N (number of examples) is large or D (number of features) is large.

   (c) TRUE / FALSE For the soft-margin SVM with decision boundary $\mathbf{w}\mathbf{x} + b = 0$ where $\mathbf{w} = (1, 2)$ and $b = 3$, the example $(\mathbf{x}, y) = ((4, 5), -1)$ has hinge loss 18.

   (d) TRUE / FALSE For training data $\{(\mathbf{x}, y)\}$ such that $\mathbf{x}_i \neq \mathbf{x}_j$ for all i and j, we can build a 3NN model that classifies the training examples without error.

   (e) TRUE / FALSE If we train a hard-margin linear SVM on linearly separable data, then discard training examples which are support vectors, and then train a new SVM on the remaining examples, the first SVM will have a wider "road" than the second.

   (f) TRUE / FALSE A linear SVM with decision boundary $(1, 2, 2) \cdot \mathbf{x} - 2 = 0$ has a smaller margin between $+1$ and -1 support vectors than one with boundary $(1, 4, 8) \cdot \mathbf{x} + 3 = 0$.

   (g) TRUE / FALSE Every decision tree regression function is a step function.
   Hint: A *step function* is a function that is constant over each of one or more intervals.

   (h) TRUE / FALSE Every k-NN regression function is a step function.
   Hint: A *step function* is a function that is constant over each of one or more intervals.

   (i) TRUE / FALSE $1 + 1 = 2$. (This is not a trick question.)

   (j) TRUE / FALSE Gradient descent can fail to converge to a global minimum if it gets stuck in a local minimum.

6. Consider the gradient descent algorithm.

   (a) Consider applying gradient descent with step size $\alpha = 0.1$ to find the $\mathbf{x}$ that minimizes the function $f(\mathbf{x}) = f\left((x^{(1)}, x^{(2)})\right) = \left(x^{(1)} - 1\right)^2 + \left(x^{(2)} - 2\right)^2$ starting from $\mathbf{x}_0 = (0, 0)$. Find the value $\mathbf{x}_1$ after one iteration.

   (b) Mark each statement as TRUE or FALSE.

      i. _____ Gradient descent can fail to converge on a convex function if step size α is such that it gets stuck in a cycle, oscillating between two or several values.

      ii. _____ For a non-convex function, gradient descent can fail to converge by descending without bound.

      iii. _____ Gradient descent can fail to converge on a convex function if it gets stuck in a local minimum.

      iv. _____ Gradient descent can fail to converge on a convex function if the step size $\alpha > 0$ is too small.

7. Here are some questions about decision trees.

(a) Consider a classification decision tree node containing the set of examples $S = \{(\mathbf{x}, y)\}$ where $\mathbf{x} = (x_1, x_2, x_3)$:

S

x_1	x_2	x_3	y
2	11	12	0
1	6	14	0
0	8	17	1
3	10	15	0
4	7	16	1
5	9	13	0

i. The entropy of this node in bits is _____.

ii. The (feature, threshold) pair (j, t) that yields the best split for this node is feature $j =$ _____ and threshold $t =$ _____.

(b) Consider a regression decision tree with `max_depth=1` (that is, the root node is split once into two leaves) made from the set of examples $S = \{(\mathbf{x}, y)\}$ where $\mathbf{x} = x_1$:

S

x_1	y
0	10
1	11
2	12
3	13
4	23
5	24

What value does this tree predict for $x_1 = 4.5$? _____

8. Here are questions about feature engineering.

   (a) Consider the data 0, 5, 5, 5, 5, which have these summary statistics:

   - minimum 0
   - mean 4
   - median 5
   - maximum 5
   - (population) standard deviation 2

   Do standardization rescaling on feature `x`:

(input) `x`	(output) `x_rescaled`
0	
5	
5	
5	
5	

   (b) Use one-hot encoding to transform the categorical feature `activity` into binary features with reasonable names.

(input) `activity`	(output)
swim	
bike	
run	
bike	

9. In linear regression we minimize the *mean squared error* (MSE).

   (a) Find the MSE for the points $(0, 0)$ and $(1, 2)$ relative to the line $\hat{y} = f_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w}\mathbf{x} + b$, where $\mathbf{w} = 2$ and $b = 3$.

   (b) For the best-fitting line for these data, $\mathbf{w} =$ _____ and $b =$ _____.