CPSC 340, 2017W1: SOLUTIONS TO FINAL EXAM

Question 1. **(12 points)**

Answer the questions below. *Be concise: avoid spending valuable time on lengthy answers.*

2 pts (a) Describe a setting where using a validation set to choose hyper-parameters can lead to overfitting?

> **Solution:** You try out a large of number of models on the validation set, and choose the one that performs the best.

2 pts (b) What is the difference between using a validation set and using cross-validation?

> **Solution:** A validation set is part of your training data that is held out during training to validate the performance of a trained model. In cross-validation you partition your examples and validate on each partition (training on the remaining partitions).

2 pts (c) Why is the IID assumption important in supervised learning?

> **Solution:** Under this assumption, we expect our test data to behave similarly to the training data.

2 pts (d) Consider a variation on decision stumps where each stump can consider k different features instead of just one. How would k affect the two parts of the fundamental trade-off?

> **Solution:** As k increase our training error will go down but our approximation error will go up.

2 pts (e) Consider the following variation on the "condensed" k-nearest neighbour algorithm from your assignment: we make our predictions based on the m "best" training examples rather than all n training examples (for a fixed m). Would this be a parametric or a non-parametric model?

> **Solution:** It's parametric. (We're no longer storing all n examples but are only storing a fixed number m.)

2 pts (a) Supervised learning models take in an x_i and output a y_i, while many clustering methods also learn to predict a y_i given an x_i. What is the key difference in their training?

> **Solution:** In supervised learning we're giving the y_i values during training.

Question 2. **(12 points)**

Answer the questions below. *Be concise: avoid spending valuable time on lengthy answers.*

2 pts (a) Consider an ensemble clustering method that generates m different boostraps of the data. It then fits a k-means model (with a random initialization) to each of the boostraps. To form the final clustering for example x_i, it chooses the y_i that is most common across the m clusterings. Would this be an effective or an inneffective ensemble method?

Solution: Beacuse of label switching. (The actual y_i values are meaningless.)

2 pts (b) Suppose we have a supervised learning problem where we think the examples x_i form clusters. To deal with this, we combine our training and test data together and fit a k-means classifier. We then add the cluster number as an extra feature, fit our supervised learning model based on the training data, then evaluate it on the test data. What have we done wrong?

Solution: We violated the golden rule, as the test data has now influenced the training procedure.

2 pts (c) What is the advantage of trying k-means with several different initializations? Should we run DBSCAN (density-based clustering) with several different initializations?

Solution: We might be able to find a better local minimum after multiple initializations. DBSCAN is not sensitive to initialization so we shouldn't use different initializations.

2 pts (d) What is an advantage of using hierarchical clustering over "flat" clustering methods like k-means and DBSCAN.

Solution: Hierarchical clustering gives more information. Or hierarchical clustering gives clusterings at multiple scales.

2 pts (e) How does λ in an L0-regularizer (like BIC) affect the sparsity pattern and the two parts of the fundamental trade-off?

Solution: As λ increase the training error goes up, the sparsity goes up, and the approximation error goes down.

2 pts (f) Instead of the L1-regularizer $\lambda \sum_{j=1}^{d} |w_j|$, consider a regularizer of the form $\lambda \sum_{j=1}^{d} \sqrt{|w_j|}$. What would an advantage and disadvantage of this regularizer be compared to L1-regularization? (Hint: think about the difference between the L2-regularizer and the L1-regularizer.)

Solution: It gives sparser solutions but it's non-convex.

Question 3. (12 points)

Answer the questions below. *Be concise: avoid spending valuable time on lengthy answers.*

2 pts (a) In Part 3 of the course, why did we add a column of "1" values to our matrix X? Does it make sense to use this trick for the methods from Part 1 of the course (decision trees, naive Bayes, k-nearest neighbours, etc.)?

Solution: In part 3 we added it so that the linear model isn't forced to go through 0 (it can have a non-zero y-intercept). We don't need this for the methods from Part 1. (They don't have a "default" value when $x_i = 0$.)

2 pts (b) For supervised training of a linear model $w^T x_i$ with $y_i \in \{-1, +1\}$, why do we use the logistic loss intead of the squared error?

Solution: The squared error penalizes for being "too right".

2 pts (c) What is the key difference between "one vs. all" logistic regression and training using the softmax loss?

Solution: We train all of our parameters simulateneously (to encourage the maximum value of $w_c^T x_i$ to be $w_{y_i}^T x_i$).

2 pts (d) Give a supervised learning scenario where you would use the Laplace likelihood and a scenario where you would use a Laplace prior.

Solution: You have outliers (Laplace likelihood). You want a sparse solution (Laplace prior).

2 pts (e) What are 3 uses for latent-factor models?

Solution: Possible answers are visualization, dimensionality reduction, features for supervised learning, discoverying/interpreting factors, outlier detection.

2 pts (f) What are two reasons that the minimizer W of the PCA objective is not unique?

Solution: Possible answers are scaling, sign changes, rotation, label switching.

Question 4. (12 points)

Answer the questions below. *Be concise: avoid spending valuable time on lengthy answers.*

2 pts (a) We said that PCA and k-means are trying to optimize the same objective function (sum of squared errors)? If we could find the optimal k-means clustering, why would it have a higher value of this objective function than PCA?

Solution: PCA allows arbitrary Z while k-means constrains Z to have at most one non-zero per row.

2 pts (b) In what setting would we prefer to use ISOMAP over PCA?

Solution: The data lies on a non-linear manifold.

2 pts (c) What do we use the backpropagation algorithm for?

Solution: Computing the gradient of a neural network.

2 pts (d) Consider a deep neural network with 1 million hidden units. Is this a parametric or a non-parametric classifier?

Solution: Parametric.

2 pts (e) A Coursera course says that convolutional neural networks can now achieve the best test error on every learning problem. Which theorem we discussed in class proves that this is non-sense?

Solution: No free lunch.

2 pts (f) What are two reasons that convolutional neural networks overfit less than classic neural networks?

Solution: Possible answers are sparsity in the weights, repeated values in the weights, max pooling.

Question 5. **(11 points)**

Match each loss function and regularizer to the corresponding property by writing the corresponding number next to each item:

1 pt (a) 0-1 loss.

1 pt (b) Absolute loss.

1 pt (c) Hinge loss.

1 pt (d) Squared loss.

1 pt (e) Softmax loss.

1 pt (f) L0-regularization.

1 pt (g) L1-regularization.

1 pt (h) L2-regularization.

1 pt (i) L∞-regularization.

1 pt (j) Logistic loss.

1 pt (k) Max of absolute residuals loss.

1. Gaussian likelihood.
2. Makes least squares solution unique.
3. Makes variables have the same magnitude.

4. Non-convex feature selection.
5. Number of training errors.
6. Regularization and feature selection.
7. Robust to outliers.
8. Smooth loss for binary classification.
9. Training all w_c simultaneously.
10. Tries to get the outliers right.
11. Upper bound on 0-1 loss.

Solution: (a) 5. (b) 7. (c) 11. (d) 1. (e) 9. (f) 4. (g) 6. (h) 2. (i) 3. (j) 8. (k) 10.

Question 6. **(10 points)**

Consider the dataset below, which has 10 training examples and 2 features:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}.$$

Suppose you want to classify the following test example:

$$\hat{x} = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

2 pts (a) What value of y_i would a 4-nearest neighbour classifier predict for this test example?

Solution: It would predict +1 (three out of four neighbours are +1).

2 pts (b) What value of y_i would a logistic regression model predict if it had the below weights?

$$w = \begin{bmatrix} 0.75 \\ -0.80 \end{bmatrix}$$

Solution: It would predict -1 (because $1 \cdot .75 + 1 \cdot (-.8) = -.05$)

3 pts (c) Below are the probabilities needed in a naive Bayes classifier to make a decision on the test example. Compute the missing values "?" values.

- $p(y = 1) = 3/5$
- $p(y = -1) =?$
- $p(x_1 = 1|y = 1) = 2/3$
- $p(x_2 = 1|y = 1) =?$
- $p(x_1 = 1|y = -1) =?$
- $p(x_2 = 1|y = -1) =?$

Solution:

- $p(y = -1) = 2/5$
- $p(x_2 = 1|y = 1) = 1/2$
- $p(x_1 = 1|y = -1) = 3/4$
- $p(x_2 = 1|y = -1) = 1/2$

3 pts (d) Under the naive Bayes model and your estimates of the above probabilities, what is the most likely label for the test example? (Show your work.)

Solution:

$$\begin{aligned}
p(y = 1|x_i) &\propto p(x_1 = 1|y = 1)p(x_2 = 1|y = 1)p(y = 1)\\
&= (2/3)(1/2)(3/5)\\
&= 6/30\\
&= 2/10.
\end{aligned}$$

$$\begin{aligned}
p(y = -1|x_i) &\propto p(x_1 = 1|y = -1)p(x_2 = 1|y = -1)p(y = -1)\\
&= (3/4)(1/2)(2/5)\\
&= 6/40\\
&= 1.5/10.
\end{aligned}$$

It would predict $y = 1$ (the values might be different depending on what they put in the previous question).

Question 7. **(6 points)**

Let n be the number of training examples, d be the number of features, and T be an upper bound on the number of iterations of gradient descent iterations we perform. Recall that the logistic regression model minimizes an objective of the form

$$f(w) = \sum_{i=1}^{n} \log(1 + \exp(-y_i w^T x_i)),$$

and its gradient has the form

$$\nabla f(w) = -\sum_{i=1}^{n} \frac{y_i}{1 + \exp(y_i w^T x_i)} x_i.$$

$\boxed{\text{3 pts}}$ (a) What is the cost of training a logistic regression model on the training set? (State your reasoning.)

Solution: For both the function and gradient, for each of the n training examples the bottleneck is computing $w^T x_i$ which costs $O(d)$ (scaling each element of x_i in the gradient has the same cost). We need to do this once for each gradient descent iteration giving a cost of $O(ndT)$.

$\boxed{\text{3 pts}}$ (b) What is the cost using forward selection to try to select the features that optimize a logistic regression model with an L0-regularization penalty like BIC? (State your reasoning.)

Solution: Each "iteration" of forward selection will fit $O(d)$ logistic regression models, and we can do at most d such "iterations". This gives a cost of $O(d^2)$ times whatever is put as the cost of fitting one model above. If they give the correct answer for the previous question this would give $O(nd^3T)$.

Question 8. **(8 points)**

$\boxed{\text{4 pts}}$ (a) Consider a variation on L2-regularized least squares where we have weights on the training examples and within the regularizer, with an objective of the form

$$f(w) = \frac{1}{2}(Xw - y)^T Z(Xw - y) + \frac{\lambda}{2}\|Aw\|^2.$$

Here, Z is an $n \times n$ diagonal matrix with non-negative values z_i along the diagonals and A is a $k \times d$ matrix. Write a linear system whose solution gives the minimum of this (convex quadratic) objective function.

Solution: We can expand this to give

$$\frac{1}{2}w^T X^T Z X w + w^T X^T Z y + \frac{1}{2}y^T Z y + \frac{\lambda}{2}w^T A^T A w.$$

The gradient is given by

$$\nabla f(w) = X^T Z X w - X^T Z y + \lambda A^T A w.$$

Setting this equal to zero and factorizing out w gives

$$(X^T Z X + \lambda A^T A)w = X^T Z y.$$

The minimizer is any solution to this linear system.

4 pts

(b) Consider a function $P(r)$ that takes an $n \times 1$ vector r and returns an $n \times 1$ vector u, where $u_i = r_i$ if r_i is positive and $u_i = 0$ otherwise. Use this function to write the hinge loss with weighted L1-regularization,

$$f(w) = \sum_{i=1}^{n} \max\{0, 1 - y_i w^T x_i\} + \lambda \sum_{j=1}^{d} v_j |w_j|,$$

in matrix and norm notation. You may find it helpful to use Y as a diagonal matrix with the y_i along the diagonal, V as a diagonal matrix withthe non-negative values v_j along the diagonal, and 1 as a vector containing all ones.

Solution:

$$f(w) = \|P(1 - YXw)\|_1 + \lambda\|Vw\|_1.$$

Another possibility is

$$f(w) = 1^T P(1 - YXw) + \lambda\|Vw\|_1.$$

Question 9. **(9 points)**

3 pts

(a) Consider an L2-regularized multi-class objective function, where for each training example we have a set $\mathcal{C}$ of pairs of classes (c_1, c_2) where we prefer class c_1 to c_2,

$$f(w) = \sum_{i=1}^{n} \sum_{(c_1,c_2)\in\mathcal{C}} \max\{0, 1 - w_{c_1}^T x_i + w_{c_2}^T x_i\} + \lambda\|W\|_F^2.$$

Show that this function is convex (you can assume $\lambda > 0$).

Solution: The second term is convex because squared norms are convex, and convexity is preserved until multiplication by positive values (in this case λ.

The first term is convex because it's a sum of convex terms. Specifically, each max is a sum of a constant function (which must be convex) and a linear function (which must be convex), and the max of convex functions is a convex function.

In PCA we approximate a centered matrix X by the matrix product ZW. What are the sizes of the following quantities we used in describing the PCA model?

1 pt

(a) z_i

Solution: $k \times 1$

1 pt

(b) z_{ic}

Solution: 1×1 (scalar)

1 pt (c) w_j

Solution: $k \times 1$

1 pt (d) W_c

Solution: $1 \times d$ (saying $d \times 1$ is also ok)

1 pt (e) $w_j^T z_i$

Solution: 1×1 (scalar)

1 pt (f) Wx_i

Solution: $k \times 1$

Question 10. **(7 points)**

Consider a supervised learning problem where we have training examples X with associated labels y, and test examples $\hat{X}$. Let Z denote a matrix obtained from a change of basis of X, and similarly let $\hat{Z}$ be the same change of basis applied to $\hat{X}$.

Let n be the number of training examples, t be the number of test examples, d be the number of features, and k be the number of features after the change of basis from X to Z. What is the cost in $O()$ notation of the following operations in terms of n, d, t, and k?

3 pts (a) Fitting an L2-regularized least squares model on the training data $\{X, y\}$ and then applying it to the test data $\hat{X}$,

$$\hat{y} = \hat{X}w, \quad \text{where} \quad w = (X^TX + \lambda I)^{-1}X^Ty.$$

Solution: $O(nd^2 + d^3)$ for training and $O(td)$ for testing, giving $O(nd^2 + d^3 + td)$.

2 pts (b) Performing the same two operations under a change of basis (you can assume that forming each element z_{ij} of Z costs $O(1)$, and forming an element of $\hat{Z}$ has the same cost),

$$\hat{y} = \hat{Z}w, \quad \text{where} \quad w = (Z^TZ + \lambda I)^{-1}Z^Ty.$$

Solution: $O(nk^2 + k^3)$ for training and $O(tk)$ for testing, giving $O(nk^2 + k^3 + tk)$.

2 pts (c) Computing the same prediction using the "other" normal equations,

$$\hat{y} = \hat{Z}w, \quad \text{where} \quad w = Z^T(ZZ^T + \lambda I)^{-1}y.$$

Solution: $O(n^2k + n^3)$ for training and $O(tk)$ for testing, giving $O(n^2k + n^3 + tk)$.

Question 11. **(7 points)**

4 pts (a) The student-t likelihood has the form

$$p(y_i|x_i, w) = \frac{\Gamma\left(\frac{\eta+1}{2}\right)}{\sqrt{\eta\pi}\Gamma\left(\frac{\eta}{2}\right)}\left(1 + \frac{(w^Tx_i - y_i)^2}{\eta}\right)^{-\frac{\eta+1}{2}},$$

where η is the "degrees of freedom" (which is at least 1) and Γ is the "gamma" function. If we use an independent Laplace prior with a mean of 0 and scale of $1/\lambda$ for each w_j,

$$p(w_j) = \frac{\lambda}{2}\exp(-\lambda|w_j|),$$

derive an objective function that is equivalent to performing MAP estimation on n IID training examples in the model (you should simplify as much as possible, but don't have to use matrix/norm notation).

Solution:

$$f(w) = \sum_{i=1}^n \log\left(1 + \frac{(w^Tx_i - y_i)^2}{\eta}\right) + \lambda\sum_{j=1}^d |w_j|.$$

It's fine if they try to put it in matrix notation, as long as they define what terms they use. It's also fine if they include the $(\eta+1)/2$ factor on the first term.

3 pts (b) State whether the above objective function has or does not have each of the below 3 properties.

- Robust to outliers.
- Sets w_j values to exactly 0.
- Is convex.

Solution:

- It's robust to outliers.
- It's gives sparse w_j.
- It's not convex.