CS534 — Midterm

Name:

1. (**Short questions**)

   a. (5pt) Define the perceptron loss function that is optimized by the Perceptron algorithm.
   Hinge loss:

$$J(w) = \frac{1}{N}\sum_{i=1}^{N} \max(0, -y_i w \cdot x_i)$$

   b. (10pt) Consider a binary random variable Z with $P(Z = 1) = p$. To estimate p, we sample Z n times, resulting in n observations of Z denoted by Z_i, $i = 1, \cdots, n$. Write down the log-likelihood function of the parameter p and derive the maximum likelihood estimation of p.

$$l(p) = \log \prod_{i=1}^{n} p^{Z_i}(1-p)^{1-Z_i}$$

$$= \sum_{i=1}^{n} [Z_i \log p + (1 - Z_i)\log(1-p)] = \log p \sum_i Z_i + \log(1-p)\sum_i (1 - Z_i)$$

$$= n_1 \log p + n_0 \log(1-p)$$

$$\frac{dl}{dp} = \frac{n_1}{p} - \frac{n_0}{1-p} = 0 \Rightarrow$$

$$n_1 - n_1 p = n_0 p \Rightarrow$$

$$p = \frac{n_1}{n_0 + n_1}$$

c. (5pt) Voted Perceptron stores all intermediate linear separators ($\mathbf{w}$'s) and takes a weighted vote as described below:

$$h(\mathbf{x}) = sign(\sum_{i=1}^{N} c_i \cdot sign(\mathbf{w}_i^T \mathbf{x}))$$

Please explain what is c_i (use your words, not pseudocode).

Answer: *c_i is the total number of examples that w_i correctly classifies before making a mistake. It provides a rough estimate of the "goodness" of w_i.*

d. (10pt) What probabilistic assumptions do Linear discriminant analysis and Logistic regression make respectively? Which of these two methods makes stronger assumption?

Answer: *LDA assumes that the data of each class follows a Gaussian distribution. The means are different but the covariance matrix is shared by both classes. Logistic Regression assumes that $p(y|x) = \frac{1}{1+exp(-w^T x)}$. Although the modeling assumption of LDA also leads to $p(y|x) = \frac{1}{1+exp(-w^T x)}$, it makes stronger assumption because it assumes specific (Gaussian) distribution for $p(x|y)$ while Logistic regression does not.*

e. (6pts) What is the Naive Bayes assumption? Consider the following data set with two input features (*temperature* and *season*). Is the naive bayes assumption satisfied for this problem?

Temperature	Season	Electricity Usage (Class)
Below Average	Winter	High
Above Average	Winter	Low
Below Average	Summer	Low
Above Average	Summer	High

Answer: *Naive Bayes assumption assumes that the features are independent given the class label. For this problem, the assumption does not hold. Because for electricity usage to be high, if S=winter, the temperature needs to be below average, whereas if S=summer, then temperature needs to be above average. They are clearly correlated given class label.*

2. (12pts) **(Issue of overfitting)** How will the following changes impact overfitting (decrease, increase or no change):

   – Pruning decision tree.
   Answer: *Decrease. Pruning decision tree will lead to simpler hypothesis and reduce overfitting.*

   – Changing k from 5 to 1 for K-Nearest Neighbor classifier.
   Answer: *Increase. Smaller k leads to more complex decision boundary, thus more overfitting.*

   – Decreasing λ for the regularized linear regression $\frac{1}{2}\sum_i(y_i - \mathbf{w^T x_i})^\mathbf{2} + \lambda|\mathbf{w}|^\mathbf{2}$.
   Answer: *Increase. Smaller regularization strength allows the weight coefficients to be larger, thus permits more overfitting.*

   – Increasing c for soft margin SVM objective function $min_{\mathbf{w,b}}|\mathbf{w}|^\mathbf{2} + \mathbf{c}\sum_\mathbf{i}\xi_\mathbf{i}$
   Answer: *Increase. Larger c will charge heavier penalty for the slack variable (which can be viewed as "error", it is in fact the commonly used "hinge loss".), thus fits data more closely, leading to more overfitting.*

   – Increasing the training data size.
   Answer: *More training data will reduce the chance of fitting to peculiarities of the training data (that occurs due to random chance.)*

   – Changing the output layer activation function of neural net from the logistic function to the identity function.
   Answer: *This question turns out to be more complicated than originally thought. On the surface, the identity function, which is linear, is less complex than the nonlinear logistic function, thus potentially leads to less complex hypothesis, suggesting less overfitting. On the other hand, the total number of parameters remains the same and suggests similar complexity. As such, this problem is removed from the exam.*

3. **(Naive Bayes)** Consider the following training data set.

X_1	X_2	X_3	Y
1	1	1	0
1	1	0	0
0	0	0	0
0	1	0	1
1	0	1	1
0	1	1	1

a. [10pt] Please apply the Naive Bayes classifier without Laplace smoothing to this data set and compute $P(y = 1|X)$ for $X = (1, 0, 0)$. Show your work.

Answer:

$$p(y = 1|X) = \frac{p(1, 0, 0|y = 1)p(y)}{p(X)} = \frac{\frac{1}{3}\frac{1}{3}\frac{1}{3}\frac{1}{2}}{p(X)} = \frac{1}{54 * P(X)}$$

$$p(y = 0|X) = \frac{p(1, 0, 0|y = 0)p(y)}{p(X)} = \frac{\frac{2}{3}\frac{1}{3}\frac{2}{3}\frac{1}{2}}{p(X)} = \frac{4}{54 * P(X)}$$

Note that the above two needs to add up to 1, thus we have $p(y = 1|X) = 1/5$.

b. [5pt] Given the following cost matrix, what is the optimal prediction for the given X? (Note that Y is the true label and $\hat{Y}$ is the predicted label.)

	$Y = 1$	$Y = 0$
$\hat{Y}$=1	0	1
$\hat{Y}$=0	5	0

Answer:

Expected loss for $\hat{y} = 1$*:* $0 * P(y = 1|X) + 1 * P(y = 0|X) = 4/5$
Expected loss for $\hat{y} = 0$*:* $5 * P(y = 1|X) + 0 * P(y = 0|X) = 5/5 = 1$
Thus we predict $\hat{y} = 1$ *because it minimizes expected loss.*

4 **(SVM)**

a. (6pts) Provide the definitions of functional margin and geometric margin. Explain why functional margin is not a proper objective to minimize.
Answer: *Functional margin:* $y(w^T x + b)$
Geometric margin: $\frac{y(w^T x+b)}{|w|}$
Functional margin is not appropriate as an objective function because it can be scaled arbitrarily without changing the decision boundary.

b. (6pts) What is the kernel trick? What is the advantage of using the kernel trick compared to explicitly mapping the features to a nonlinear dimensional space? ***Answer:*** *The kernel trick replaces the linear dot product with a kernel function, which is equivalent to mapping the features to a higher dimensional space then taking the dot product in the mapped space. This has significant computational advantage over explicitly mapping the features then taking dot product.*

5 **(Learning theory)**

a. (5pts) Consider the following upper bound on the generalization error of the learned hypothesis (which minimizes the training error). Please use this bound to explain how the complexity of H may influence the generalization error of the learned hypothesis.

$$\epsilon(h_L) \leq \epsilon(h^*) + 2\sqrt{\frac{1}{2m}(\ln|H| + \ln\frac{1}{\delta})}$$

Answer: *As we increase the complexity of the hypothesis space (thus increasing $|H|$), it influences the upper bound in two ways. For the first term, it will likely contain an $h^* \in H$ that achieves lower $\epsilon(h^*)$. On the other hand, the second term will increase as $|H|$ increases. Minimizing the upper bound will require trading off these two aspects to find the proper hypothesis space.*

b. (8pts) Consider the hypothesis space H defined by circles in a 2-d space (note that the region inside the circle is positive). Prove that VC dimension of this hypothesis space ≥ 3. What do you need to show to prove that the VC dimension is exactly 3?

Answer: *To prove that $VC(H) \geq 3$, we just need to provide one set of 3 points and show that H could shatter them, i.e., H contains an h that will correctly classify all instances no matter how we label them. This is rather simple. Three points that do not fall on a line will suffice. I will omit the figures showing the labeling here. To prove that $VC(H) = 4$, we will need to show that for any set of 4 points, H cannot possibly shatter the set. This is trickier to prove. The basic idea is to group all possibilities into general cases, and show for each case this is not possible.*