Discussion 6B Solutions

1 Maximum Likelihood Estimation

Recall that a Geometric distribution is a defined as the number of Bernoulli trials needed to get one success. $P(X = k) = p(1-p)^{k-1}$.
We observe the following samples from a Geometric distribution:
$x_1 = 5$, $x_2 = 8$, $x_3 = 3$, $x_4 = 5$, $x_5 = 7$
What is the maximum likelihood estimate for p?

$$\begin{aligned}
L(p) &= P(X = x_1)P(X = x_2)P(X = x_3)P(X = x_4)P(X = x_5) && (1)\\
&= P(X = 5)P(X = 8)P(X = 3)P(X = 5)P(X = 7) && (2)\\
&= p^5(1-p)^{23} && (3)\\
\log(L(p)) &= 5\log(p) + 23\log(1-p) && (4)\\
& && (5)
\end{aligned}$$

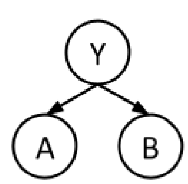
We must maximize the log-likelihood of p, so we will take the derivative, and set it to 0.

$$\begin{aligned}
0 &= \frac{5}{p} - \frac{23}{1-p} && (6)\\
p &= 5/28 && (7)
\end{aligned}$$

2 Naive Bayes

In this question, we will train a Naive Bayes classifier to predict class labels Y as a function of input features A and B. Y, A, and B are all binary variables, with domains 0 and 1. We are given 10 training points from which we will estimate our distribution.

A	1	1	1	1	0	1	0	1	1	1
B	1	0	0	1	1	1	1	0	1	1
Y	1	1	0	0	0	1	1	0	0	0



(a) What are the maximum likelihood estimates for the tables $P(Y)$, $P(A|Y)$, and $P(B|Y)$?

Y	$P(Y)$
0	3/5
1	2/5

A	Y	$P(A\|Y)$
0	0	1/6
1	0	5/6
0	1	1/4
1	1	3/4

B	Y	$P(B\|Y)$
0	0	1/3
1	0	2/3
0	1	1/4
1	1	3/4

(b) Consider a new data point ($A = 1$, $B = 1$). What label would this classifier assign to this sample?

$$
\begin{aligned}
P(Y=0, A=1, B=1) &= P(Y=0)P(A=1|Y=0)P(B=1|Y=0) && (8)\\
&= (3/5)(5/6)(2/3) && (9)\\
&= 1/3 && (10)\\
P(Y=1, A=1, B=1) &= P(Y=1)P(A=1|Y=1)P(B=1|Y=1) && (11)\\
&= (2/5)(3/4)(3/4) && (12)\\
&= 9/40 && (13)\\
& && (14)
\end{aligned}
$$

Our classifier will predict label 0.

(c) Let's use Laplace Smoothing to smooth out our distribution. Compute the new distribution for $P(A|Y)$ given Laplace Smoothing with $k = 2$.

A	Y	$P(A\|Y)$
0	0	3/10
1	0	7/10
0	1	3/8
1	1	5/8

Q3. Machine Learning: Potpourri

(a) What it the **minimum** number of parameters needed to fully model a joint distribution $P(Y, F_1, F_2, ..., F_n)$ over label Y and n features F_i? Assume binary class where each feature can possibly take on k distinct values.
$2k^n - 1$

(b) Under the **Naive Bayes assumption**, what is the **minimum** number of parameters needed to model a joint distribution $P(Y, F_1, F_2, ..., F_n)$ over label Y and n features F_i? Assume binary class where each feature can take on k distinct values. $2n(k-1)+1$

(c) You suspect that you are overfitting with your Naive Bayes with Laplace Smoothing. How would you adjust the strength k in Laplace Smoothing?

- [x] Increase k
- [ ] Decrease k

(d) While using Naive Bayes with Laplace Smoothing, increasing the strength k in Laplace Smoothing can:

- [x] Increase training error
- [ ] Decrease training error
- [x] Increase validation error
- [x] Decrease validation error

(e) It is possible for the perceptron algorithm to never terminate on a dataset that is linearly separable in its feature space.

- [ ] True
- [x] False

(f) If the perceptron algorithm terminates, then it is guaranteed to find a max-margin separating decision boundary.

- [ ] True
- [x] False

(g) In binary perceptron where the initial weight vector is $\vec{0}$, the final weight vector can be written as a linear combination of the training data feature vectors.

- [x] True
- [ ] False

(h) For binary class classification, logistic regression produces a linear decision boundary.

- [x] True
- [ ] False

(i) In the binary classification case, logistic regression is exactly equivalent to a single-layer neural network with a sigmoid activation and the cross-entropy loss function.

- [x] True
- [ ] False

(j) You train a linear classifier on 1,000 training points and discover that the training accuracy is only 50%. Which of the following, if done in isolation, has a good chance of improving your training accuracy?

- [x] Add novel features
- [ ] Train on more data

(k) You now try training a neural network but you find that the training accuracy is still very low. Which of the following, if done in isolation, has a good chance of improving your training accuracy?

- [x] Add more hidden layers
- [x] Add more units to the hidden layers