Recitation 10
SVMs, Kernels, Naïve Bayes

10-301/10-601: Introduction to Machine Learning

12/04/2020

1 Naive Bayes

Definitions:

- Bayes Rule: $p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$
- Naive Bayes Model:

  assumes attributes $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K$ are conditionally independent of each other given the class label y, and that their prediction can be written as $\hat{y} = \arg\max_y P(y|X)$, where

$$p(y|X = (\mathbf{x}_1, ..., \mathbf{x}_k)) \propto p(y)p(X|y) = p(y)\prod_{k=1}^{K} p(\mathbf{x}_k|y)$$

Questions:

1. What are some differences and similarities between Generative Models and Discriminative Classifiers?

   Generative Classifiers

   - learns the joint probability distribution $p(x, y)$
   - estimates parameters $p(X|Y), p(Y)$ from training data
   - uses Bayes rule to calculate $p(Y|X)$

   Discriminative Classifiers

   - learns the conditional probability distribution $p(Y|X)$

   Both Models

   - both are usually predicting conditional probabilities $p(Y = y|X)$

2. You love your old red mustang convertible, but you just couldn't keep up on the payments. It's now time to sell the car. To generate some interest, you decide to post an online advertisement for your car. When someone views the advertisement, you know three things - are they currently looking for a car? Do they like the color red? Do they currently have free time to look at an ad? The below table summarizes some previous data you collected. Note 1 represents "Yes" and 0 represents "No".

Clicked?	Looking?	Likes Red?	Free?
1	1	1	1
0	0	1	0
0	0	1	1
1	0	0	0
0	0	1	1
0	0	1	1
1	1	1	1
1	1	0	1
0	0	0	0

A new person visits the page. What you know is that they are looking for a car, don't like red, and currently have free time. Using a Naive Bayes approach, Do you predict that they will click on the advertisement? A table with parameter values is provided below for convenience.

feature	Clicked = 1	Clicked = 0
Looking?	3/4	0/5
Likes Red?	2/4	4/5
Free?	3/4	3/5

$$p(\text{Clicked} = 1|\text{looking} = 1, \text{likes red} = 0, \text{free} = 1) \propto p(\text{Clicked} = 1) \prod_i^k p(x_i|\text{Clicked} = 1)$$
$$= \frac{4}{9} * \frac{3}{4} * \frac{2}{4} * \frac{3}{4}$$
$$= \frac{18}{144}$$
$$= 0.125$$

$$p(\text{Clicked} = 0|\text{looking} = 1, \text{likes red} = 0, \text{free} = 1) \propto p(\text{Clicked} = 0) \prod_i^k p(x_i|\text{Clicked} = 0)$$
$$= \frac{5}{9} * \frac{0}{5} * (1 - \frac{4}{5}) * \frac{3}{5}$$
$$= 0$$

Because $0.125 > 0$, we predict that they will click.

3. Is there anything wrong here? What do we know will happen to our prediction for any person if looking = 1? Why might this be a problem? How can we fix it?

   If looking = 1, then we will always have $p(\text{Clicked} = 0|\text{looking} = 1, ...) = 0$. Therefore we will always predict that Clicked = 1. To fix it, we could add 1 to the value in every numerator and denominator to yield the following table.

feature	Clicked = 1	Clicked = 0
Looking?	4/5	1/6
Likes Red?	3/5	5/6
Free?	4/5	4/6

4. **BONUS** When Y is Boolean and $\mathbf{X} = (X_1, \cdots, X_n)$ is a vector of continuous variables, then the assumptions of the Gaussian Naive Bayes classifier imply that $P(Y \mid \mathbf{X})$ is given by the logistic function with appropriate parameters W. In particular:

$$P(Y = 1 \mid \mathbf{X}) = \frac{1}{1 + \exp(b + \sum_{i=1}^{n} w_i X_i)}$$

and

$$P(Y = 0 \mid \mathbf{X}) = \frac{\exp(b + \sum_{i=1}^{n} w_i X_i)}{1 + \exp(b + \sum_{i=1}^{n} w_i X_i)}$$

Consider instead the case where Y is Boolean and $\mathbf{X} = (X_1, \cdots, X_n)$ is a vector of Boolean variables. Prove for this case also that $P(Y \mid \mathbf{X})$ follows this same form (and hence that Logistic Regression is also the discriminative counterpart to a Naive Bayes generative classifier over Boolean features).

Hints

1. Simple notation will help. Since the X_i are Boolean variables, you need only one parameter to define $P(X_i \mid Y = y_k)$. Define $\phi_{i1} \equiv P(X_i = 1 \mid Y = 1)$, in which case $P(X_i = 0 \mid Y = 1) = (1 - \phi_{i1})$. Similarly, use ϕ_{i0} to denote $P(X_i = 1 | Y = 0)$.
2. Notice with the above notation you can represent $P(X_i \mid Y = 1)$ as follows

   $$P(X_i \mid Y = 1) = \phi_{i1}^{(X_i)}(1 - \phi_{i1})^{(1-X_i)}$$

   Note when $X_i = 1$ the second term is equal to 1 because its exponent is zero. Similarly, when $X_i = 0$ the first term is equal to 1 because its exponent is zero.

Proof:

In general Bayes Rule allows us to write

$$P(Y = 1|X) = \frac{P(Y = 1)P(X|Y = 1)}{P(Y = 1)P(X|Y = 1) + P(Y = 0)P(X|Y = 0)}$$

$$= \frac{1}{1 + \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$

$$= \frac{1}{1 + \exp(\ln(\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}))}$$

Because of our conditional independence assumption this becomes

$$\frac{1}{1+\exp(\ln(\frac{P(Y=0)}{P(Y=1)})+\sum_i \ln(\frac{P(X_i|Y=0)}{P(X_i|Y=1)}))}$$

$$=\frac{1}{1+\exp(\ln(\frac{1-\pi}{\pi})+\sum_i \ln(\frac{P(X_i|Y=0)}{P(X_i|Y=1)}))}$$

Note the final step expresses $P(Y=0)$ and $P(Y=1)$ in terms of the binomial parameter π.

Looking just at the summation term and substituting we have

$$\sum_i \ln\left(\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)=\sum_i \ln\left(\frac{\phi_{i0}^{X_i}(1-\phi_{i0})^{(1-X_i)}}{\phi_{i1}^{X_i}(1-\phi_{i1})^{(1-X_i)}}\right)$$

$$=\sum_i \ln(\phi_{i0}^{X_i}(1-\phi_{i0})^{(1-X_i)})-\ln(\phi_{i1}^{X_i}(1-\phi_{i1})^{(1-X_i)})$$

$$=\sum_i \ln(\phi_{i0}^{X_i})+\ln((1-\phi_{i0})^{(1-X_i)})-\ln(\phi_{i1}^{X_i})-\ln((1-\phi_{i1})^{(1-X_i)})$$

$$=\sum_i X_i\ln(\phi_{i0})+(1-X_i)\ln((1-\phi_{i0}))-X_i\ln(\phi_{i1})-(1-X_i)\ln((1-\phi_{i1}))$$

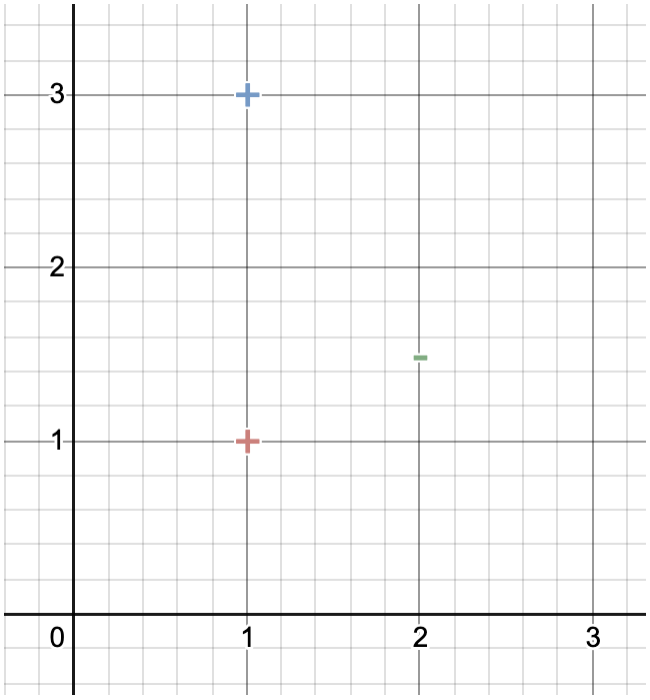
$$=\sum_i X_i\ln(\phi_{i0})+\ln(1-\phi_{i0})-X_i\ln(1-\phi_{i0})-X_i\ln(\phi_{i1})-\ln(1-\phi_{i1})+X_i\ln(1-\phi_{i1})$$

Substituting back into our equation this gives

$$\frac{1}{1+\exp(\ln(\frac{1-\pi}{\pi})+\sum_i X_i\ln(\phi_{i0})+\ln(1-\phi_{i0})-X_i\ln(1-\phi_{i0})-X_i\ln(\phi_{i1})-\ln(1-\phi_{i1})+X_i\ln(1-\phi_{i1})}$$

2 SVMs

1. What is the decision boundary and the margin if we run a Hard-Margin SVM on the following set of points?



Decision Boundary: x = 1.5, Margin = 0.5

2. A few additional data points are added to the data set in figures 2 (a) and 2 (b). Draw the new decision boundaries and give the margins corresponding to this boundaries. In which case does the decision boundary undergo a change and why?

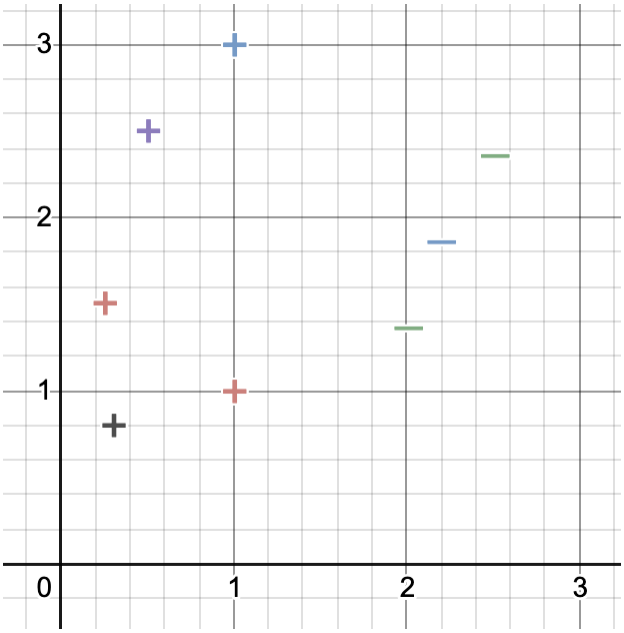


Figure 2(a)

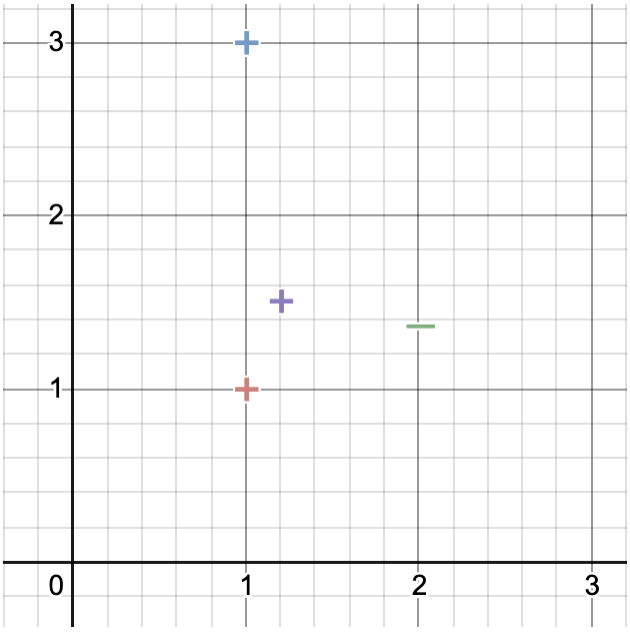
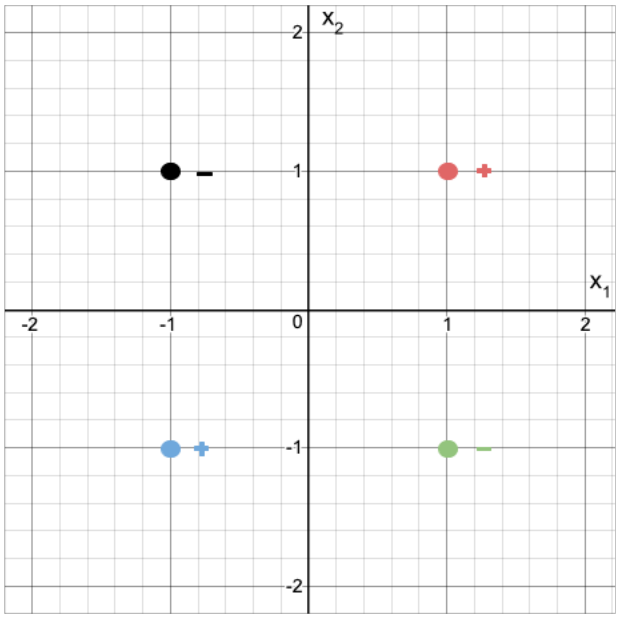


Figure 2(b)

For figure 2(a) the decision boundary remains unchanged because the support vector is the same. 2(b) there is a non-vertical decision boundary.

3 Kernels

The XOR-problem is a non-linear problem which can be represented by the plot below.

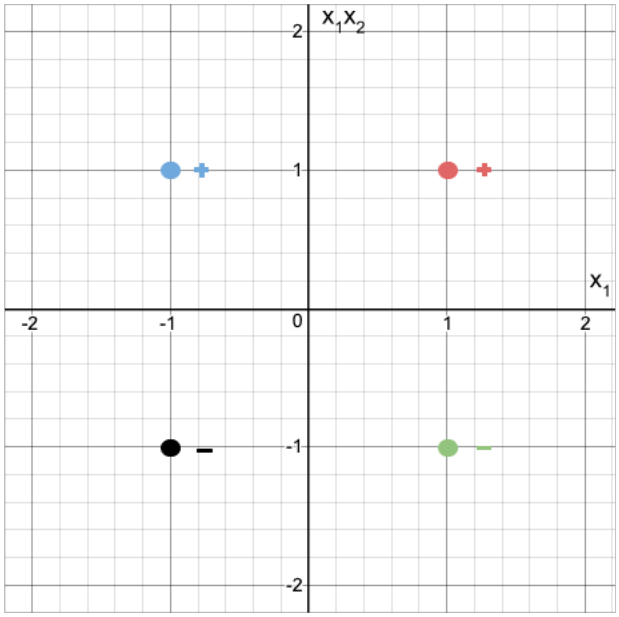


For the following questions, consider a feature transformation - $\phi([x_1, x_2]^T) = [x_1, x_1x_2]^T$

1. What is the kernel $K(x, z)$?

   $K(x, z) = \phi([x_1, x_2])^T\phi([z_1, z_2]) = [x_1, x_1x_2]^T[z_1, z_1z_2] = x_1z_1 + x_1x_2z_1z_2$

2. How is the dataset represented in the transformed space?



3. Is the dataset linearly separable in the transformed space? If so, give the boundary in the original space?

   $x_1 x_2 = 0$