582631 Introduction to Machine Learning, Fall 2016

Due December 1st–2nd. NB: Deadline for returning solutions my email (in case you can't attend a session) is 12:15 on Friday.

Problem 1 (3 + 3 + 4 + 2)

(a) (3 points) Consider a binary classification problem with $Y \in \{-1, +1\}$ and two real-valued features, $X_1, X_2 \in \mathbb{R}$. Suppose that we have learned a (Gaussian) naive Bayes classifier and obtained parameter estimates $\hat{\mu}_{-,j} = 0, \hat{\sigma}^2_{-,j} = 1$ and $\hat{\mu}_{+,j} = 0, \hat{\sigma}^2_{+,j} = 16$ for $j \in \{1, 2\}$. Further, we use a uniform class prior $\hat{p}(y) = 1/2$ for $y \in \{-1, +1\}$.

Use the Bayes formula to compute the posterior probability

$$P(Y = +1 \mid X_1 = 1, X_2 = 2)$$

(b) (3 points) Visualize the class posterior $P(Y = +1 \mid \mathbf{x})$, where $\mathbf{x} = (x_1, x_2)$ contains the observed feature values on a suitable grid of points.

Hint: This is quite similar to Exercise 3.1.

(c) (4 points) (This may look quite difficult, but in case you have any trouble, feel free to ask for help — we've provided plenty of opportunities for doing that)

Show that the above naive Bayes classifier is actually a special case of QDA.

Hint: Both are generative classifiers with Gaussian $p(\mathbf{x} \mid y)$. What is the proper choice for the QDA parameters $\boldsymbol{\mu}$ and Σ that produces the same conditional distribution $p(\mathbf{x} \mid y)$ as the naive Bayes classifier with the above parameters?

(d) (2 points) What is the number of free parameters that are needed to specify a (Gaussian) naive Bayes classifier for p features?

What is the number of free parameters that are needed to specify a QDA classifier?

Which model is more complex? What do you think this will mean in terms of overfitting and the number of training data points required to achieve the asymptotic error (recall Exercise 3.3) of each classifier?

Hint: The number of free parameters is the minimum number of values that one needs to learn from the data. For example, while a covariance matrix has $p \times p = p^2$ elements, the number of *free* parameters required to specify the covariance matrix is $p(p+1)/2 < p^2$ because the matrix is symmetric.

(Exercises continued on the next page...)

Problem 2 (2 + 2 + 2 + 2 points)

Assume that we have a situation where the class variable Y can take three values $\{0, 1, 2\}$, and that there are two feature variables $X_1 \in \{0, 1\}$ and $X_2 \in \{0, 1, 2\}$. The true distribution from which the data is sampled is such that the class distribution $P(Y)$ is given by $P(Y = 0) = 0.4$, $P(Y = 1) = 0.3$ and $P(Y = 2) = 0.3$, and class-conditional distributions $P(\mathbf{X} \mid Y)$ are as specified below (with $\mathbf{X} = (X_1, X_2)$):

X_2 \ X_1	0	1
0	0.2	0.1
1	0.4	0.2
2	0.0	0.1

$P(X_1, X_2 \mid Y = 0)$

X_2 \ X_1	0	1
0	0.6	0.1
1	0.1	0.1
2	0.1	0.0

$P(X_1, X_2 \mid Y = 1)$

X_2 \ X_1	0	1
0	0.1	0.4
1	0.3	0.0
2	0.2	0.0

$P(X_1, X_2 \mid Y = 2)$

(a) (2 points) Let's draw some training data from the above source. First draw a single class value from $P(Y)$ as follows (if you are using R):

```
sample(0:2, size=1, replace=TRUE, prob=c(0.4,0.3,0.3))
```

This gives (by chance), $Y = 1$. Now then, we can draw the feature values from their joint distribution as follows:

```
expand.grid(0:1,0:2)[sample(1:6, 1, replace=TRUE, prob=c(0.6,0.1,0.1,0.1,0.1,0.0)),]
```

where the `prob` vector is obtained from the above conditional probability table for $Y = 1$. Lo and behold, we get $\mathbf{X} = (0, 0)$, which indeed is the most likely outcome for this class. (You may well get something else, which is the point in *random* sampling.)

Draw a training data set of $n = 100$ points using this procedure. Check that you get roughly 40 cases with $Y = 0$, and roughly 29 cases of $\mathbf{X} = (0, 0)$. (Can you see why the latter should be the case?)

(b) (2 points) Use the formula on p. 5 of the lecture slides (Lecture 7) to obtain smoothed estimates of the class conditional distributions $\hat{P}(X_i \mid Y = c)$ for each feature $i \in \{1, 2\}$ and each class $c \in \{0, 1, 2\}$. Try different smoothing parameters (maximum likelihood, Laplace, Krichesky-Trofimov).

Also, apply the same kind of smoothing to estimate the class distribution as follows

$$\hat{P}(Y = c) = \frac{n_c + \alpha}{n + 3\alpha}$$

where n_c is the number of training instances of class c, and $\alpha \in \{0, 1, 1/2\}$.

(c) (2 points) Now generate a test set of 10 000 points from the same source as the training set and apply the naive Bayes classifier you learned from the training data.

What is the test set error you obtain? (The test set error with Laplace smoothing should be between 0.4–0.6.)

Repeat with training sets of size $n = 25, 50, 100, 200, 400, 800, 1600, 3200, 6400$ and plot the test set error as a function of the training set size. (The asymptotic error is 0.4.) Does the smoothing method have an effect on the error?

(d) (2 points) Apply logistic regression to the same data and compare the results to those of the naive Bayes classifier. Analyse and try to explain the results. You may ask whether the assumption underlying the naive Bayes classifier is valid?

Hint: For logistic regression, use the function `multinom` from package `nnet`, which works very similar to the `glm` function but allows more than two classes; about `glm` and logistic regression, see the Lab on pp. 156–161 of the textbook. You should treat the covariates as qualitative (see Sec. 3.3.1 of the textbook), which is achived by converting the columns to `factor`s as follows:

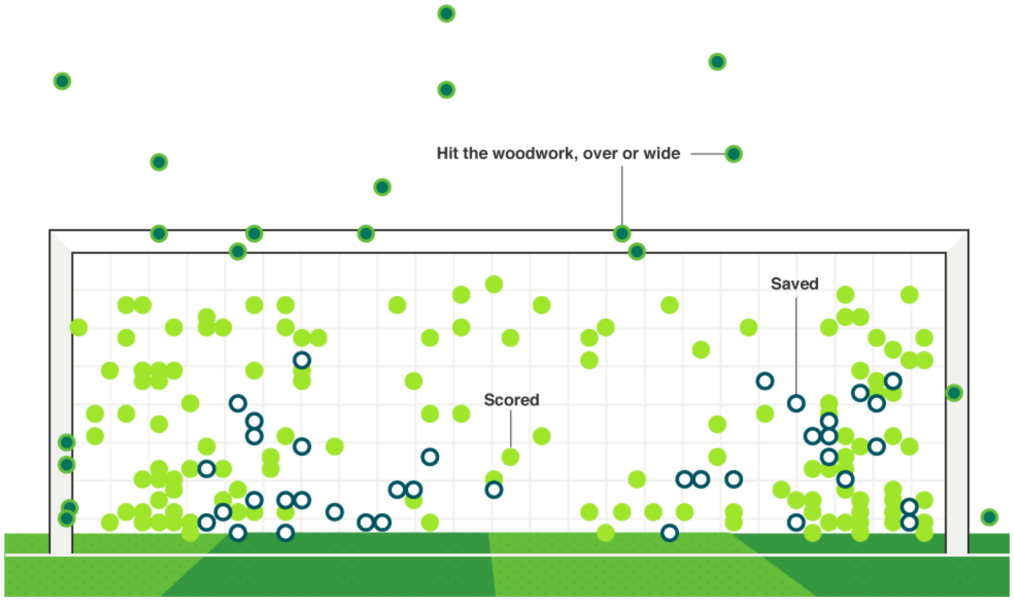
```
D <- data.frame(Y = train$Y, X1 = factor(train$X[ ,1]), X2 = factor(train$X[ ,2]))
```

To check whether the naive Bayes assumption is valid, recall when two random variables are conditionally independent given a third random variable.

(Exercises continued on the next page...)

Problem 3 (2 + 2 points)

Consider the following infographic showing about 200 penalty shots taken at the 2010 FIFA World Cup.[1] You need only consider the light green (scored) and white (saved) shots, so that the problem becomes one of binary classification. (The dark green dots hit the woodwork — like Asamoah's — or went over or wide.)



Source: BBC, `http://www.bbc.co.uk/guides/zgg334j`

(a) (2 points) Manually sketch a decision tree to classify whether a shot will be scored or saved. The tree should make at least five splits.

(b) (2 points) Given a subset of the points, D, define its *impurity* as $Q(D) = 1 - \max_k \hat{p}_k$, where $\hat{p}_k \in [0, 1]$ denotes the proportion of the points in D that belong to class k. In other words, $Q(D)$ is the remaining fraction once the most common class is removed.

Define the impurity of a split that splits D into two subsets D_1, D_2 as

$$Q(\{ D_1, D_2 \}) = \sum_{i=1}^{2} \frac{|D_i|}{|D|} Q(D_i),$$

where $|D_i|$ denotes the number of points in set D_i. Finally, define the *gain* of the split as

$$Q(D) - Q(\{ D_1, D_2 \})$$

Use the above definitions and compute the gain of each split in your decision tree.

Hint: Don't worry if you don't count the green and white balls exactly right as long as they are "in the ballpark".

[1] You may still remember some of them: `https://www.youtube.com/watch?v=tDpx9GGH79I`.