CSE517A – Homework 2

M. Neumann

Feb 15 2018

- Please keep your written answers brief and to the point. Incorrect or rambling statements can hurt your score on a question.
- If your hand writing is not readable, we **cannot give you credit**. We recommend you type your solutions in LaTeX and compile a .pdf for each answer. **Start every problem on a new page!**
- This will be due THU **Mar 1 2018 10am** with an automatic 3-day extension. If you want your assignment graded before the midterm exam, submit **before** the actual deadline. *You will **not** be able to **update** your submission in the extension period.* Submissions that we receive in the extension period cannot be graded before the midterm.
- You may work in groups of at most 2 students.
- Submission instructions:
  - Start every problem on a **new page**.
  - Submissions will be exclusively accepted via **Gradescope**. Find instructions on how to get your Gradescope account and submit your work on the course webpage.

Problem 1 (*30 points*) Naïve Bayes

In this problem, we explore why the naïve assumption is necessary.

After learning about using Bayes rule to make predictions (without the naïve assumption), you want to apply the method to a complex problem. Keep in mind that you want to use the rule:

$$\Pr(Y \mid X) = \frac{\Pr(X \mid Y) \cdot \Pr(Y)}{\Pr(X)}$$

and you want to estimate the parameters of $\Pr(X \mid Y)$ and $\Pr(Y)$. However, before applying the method to your problem, you want to apply to a toy problem first.

(a) (*10 pts*) In the toy problem, $X = [X_1, X_2]$ (so $d = 2$), where X_i is binary. Y is also binary. You want to estimate $\Pr(X \mid Y)$ without the Naïve Bayes assumption, that is you cannot write

$$\Pr(X \mid Y) = \prod_{i=1}^{d} \Pr(X_i = x_i \mid Y = y),$$

instead, you must estimate

$$\Pr(X \mid Y) = \Pr(X_1 = x_1, \ldots, X_d = x_d \mid Y = y)$$

for all combinations of the values $x_1, \ldots, x_d$, and y. How many parameters do you have to estimate of for your toy problem?
(Here parameters refers to the estimate of $\Pr(X \mid Y) = \Pr(X_1 = x_1, \ldots, X_d = x_d \mid Y = y)$, and $\Pr(Y = y)$ for some combination of x_i's and y. In our case where $d = 2$, examples of such parameters are $\Pr(X_1 = 1, X_2 = 0 \mid Y = 0)$ and $\Pr(Y = 1)$.)

(b) (*10 pts*) After running the decision rule on your toy problem, you decide to apply it to the actual problem. However, in your problem, $d = 100$. How many parameters do you have to estimate now?

(c) (*10 pts*) When is it necessary for us to make the naïve assumption? Explain by showing how the assumption will affect one of the answers from above.

Problem 2 (*30 points*) Naïve Bayes, part II.

(a) (*10 pts*) We will attempt to write the multinomial naïve Bayes classifier's decision rule as a linear rule. Suppose that we have a document that is represented as a sequence $d = (w_1, \ldots, w_\ell)$ of length ℓ. This can be classified as:

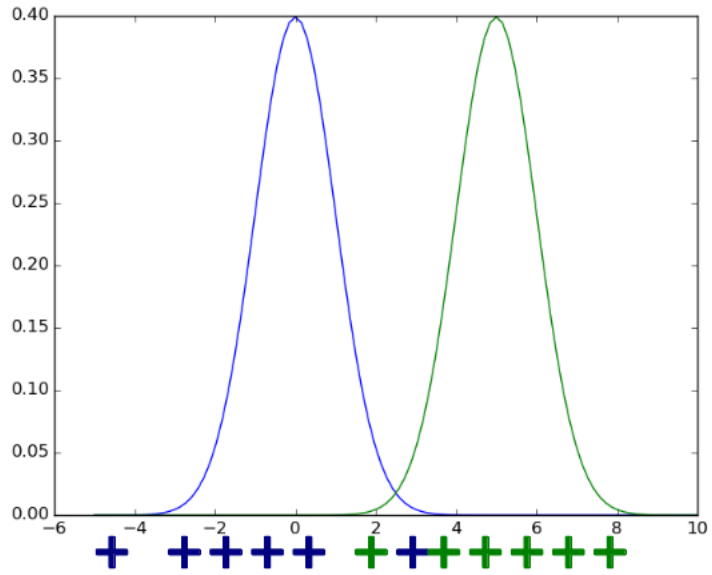
$$h(d) = \underset{y \in \{+1, -1\}}{\arg\max} \Pr(Y = y) \prod_{i=1}^{\ell} \Pr(W = w_i \mid Y = y)$$

Assuming that we have estimates for all the appropriate probabilities and none of them are zero, rewrite $h(d)$ as $h(d) = \text{sign}(\vec{v} \cdot \vec{x} + b)$. We assume that we have a vocabulary of words $\{a_1, \ldots, a_m\}$ that contains m words. Each word w_i in the document is one of the a_j, and $\vec{x}$ is a vector where each component x_j is the count of the number of times the word a_j shows up in the document. Provide $\vec{v}$, and b.

(b) (*10 pts*) Previous problems considered only those cases where X consists of discrete values. Now, we will also consider the case where X can take continuous values.

But first, (i) write down the maximum likelihood estimates (MLE) for the parameters of the naïve Bayes classifier, $\Pr(X_i \mid Y)$ and $\Pr(Y)$, where X takes discrete, categorical values.

Now let's look at naïve Bayes classifier that takes vectors of continuous values as input. In this case, we need a different formulation for $\Pr(X_i|Y)$ (we don't need to worry about $\Pr(Y)$ because Y still takes discrete values). One way is to assume that, for each discrete y, each X_i comes from a Gaussian.



For example, consider the simplified case above, where X_i takes a continuous value (a value along the x-axis) and Y can be either *blue* or *green*. As shown above, for each discrete value of Y (*blue* or *green*), X_i is a random variable from a Gaussian specific to X_i (not some other X_j) and the value of Y. As a result, we see two Gaussians, each generating *blue* data points or *green* data points.

With this, we get a Gaussian Naïve Bayes classifier. Using the assumption, (ii) write down the MLE for the parameters of $\Pr(X_i \mid Y)$ (Note: since $\Pr(X_i \mid Y = y)$ is a Gaussian its parameters are its mean $\mu_{y,i}$ and standard deviation $\sigma_{y,i}$). (iii) What is the total number of parameters?

(c) (*10 pts*) Using what we found in part (b), we will reformulate the classifier once more. Remember how a Gaussian naíve Bayes classifier is defined:

- X_i is continuous
- $\Pr(X_i \mid Y = y)$ is a Gaussian with $\mu_{y,i}, \sigma_i$ (we will assume that each σ only depends on the feature, and not the label y)
- Given the label, features are conditionally independent, just like the discrete version
- Y is a Bernoulli random variable

Now, find the weight vector $\vec{w}$ such that

$$\Pr(Y = +1|X) = \frac{\Pr(X|Y = +1) \cdot \Pr(Y = +1)}{\Pr(X)} = \frac{1}{1 + \exp[-(w_0 + \sum_{i=1}^{n} w_i X_i)]}$$

Be sure to use your answers for part (b).

Note: The result you got it precisely the formulation for logistic regression. However, Gaussian naïve Bayes and logistic regression are different learning algorithms. They output the (asymptotically) same model only under special conditions. We will explore the relation further when we cover logistic regression.

Problem 3 (*40 points*) Valid Kernels, Kernel Construction

(a) (*10 pts*) Consider $\mathbf{x}_i \in \mathbb{R}^3$ and $\phi(\mathbf{x}_i) \in \mathbb{R}^D$ with

$$\phi(\mathbf{x}_i) = [1, \sqrt{2}(\mathbf{x}_i)_1, \sqrt{2}(\mathbf{x}_i)_2, \sqrt{2}(\mathbf{x}_i)_3, (\mathbf{x}_i)_1^2, (\mathbf{x}_i)_2^2, (\mathbf{x}_i)_3^2, \sqrt{2}(\mathbf{x}_i)_1(\mathbf{x}_i)_2, \sqrt{2}(\mathbf{x}_i)_1(\mathbf{x}_i)_3, \sqrt{2}(\mathbf{x}_i)_2(\mathbf{x}_i)_3]^T,$$

compute $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. What is D? Briefly argue why considering the inner products directly instead of explicitly computing the feature mappings is more efficient.

(b) (*10 pts*)For each of the following matrices show whether it is strictly positive definite, positive semidefinite, or neither:

$$A_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 2 \end{bmatrix}.$$

(c) (*10 pts*) For positive semi-definite (psd) matrices A,B show that $A + B$ and $A \otimes B$ ($\otimes$ is the Kronecker product) are positive semi-definite.

(d) (*10 pts*) Show that the RBF Kernel corresponds to an inner product in an infinte dimensional space. (HINT: use Taylor expansion.)

Problem 4 (*30 points*) Kernelize the Perceptron Algorithm

Kernelize the perceptron algorithm for binary classification $y_i = \{+1, -1\}$.

(0) Initialize $\vec{w} = 0$

REPEAT until convergence:

(1) Pick $(\vec{x}_i, y_i)$ randomly from D.

(2) If $y_i \vec{w}^T \vec{x}_i \leq 0$ then make the update $\vec{w} \leftarrow \vec{w} + y_i \vec{x}_i$, otherwise do nothing.

(a) (*10 pts*) Show that $\vec{w}$ can be written as a linear combination of the training inputs.

(b) (*10 pts*) Derive the kernelized classifier $h(\vec{x})$.

(c) (*10 pts*) State the kernelized version of the learning algorithm.