Problem 2: Gaussian Naive Bayes

In class it was briefly mentioned that the Naive Bayes Classifier can be extended to predictor variables with continuous values (instead of just binary variables). We will derive such an approach here

Consider a data set where each example $(\vec{x}, y)$ contains a data vector $\vec{x} \in \mathbb{R}^d$ and a label $y \in \{0, 1\}$. Each y is modeled a Bernoulli random variable, which equals 1 with probability p and 0 with probability $1 - p$. To model $\vec{x}$ we have two lists of mean/variances pairs:

$$(\mu_{0,1}, \sigma^2_{0,1}), (\mu_{0,2}, \sigma^2_{0,2}), \ldots, (\mu_{0,d}, \sigma^2_{0,d}) \text{ and } (\mu_{1,1}, \sigma^2_{1,1}), (\mu_{1,2}, \sigma^2_{1,2}), \ldots, (\mu_{1,d}, \sigma^2_{1,d}). \quad (1)$$

If y equals 0, then the j^{th} entry of $\vec{x}$ is modeled as an *independent* Gaussian (normal) random variable with mean $\mu_{0,j}$ and variance $\sigma^2_{0,j}$. Alternatively, if y equals 1, then the j^{th} entry of $\vec{x}$ is modeled as an *independent* Gaussian random variable with mean $\mu_{1,j}$ and variance $\sigma^2_{1,j}$.

(a) Given a training data set $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$ write down expressions for estimating all model parameters $\mu_{i,j}$ and $\sigma^2_{i,j}$ from the data.

Solution: First set p to be the fraction of all training data with label 1. Next let $S_0 \subseteq \{1, \ldots, n\}$ contain all indices k such that $y_k = 0$ and let $S_1 \subseteq 1\{1, \ldots, n\}$ contain all indices k such that $y_k = 1$. For each $i \in \{0, 1\}$ and $j \in 1, d$, let $\mu_{i,j} = \frac{1}{|S_i|} \sum_{k \in S_i} \vec{x}_k[j]$. Let $\sigma^2_{i,j} = \frac{1}{|S_i|} \sum_{k \in S_i} (\vec{x}_k[j] - \mu_{i,j})^2$.

(b) Given a new unlabled predictor vector $\vec{x}_{new}$ we would like to predict class label y_{new} using a *maximum a posterior* (MAP) estimate. In other words, we want to choose y_{new} to maximize the posterior probability $p(y_{new} \mid \vec{x}_{new})$. Write down an expression for $p(y_{new} \mid \vec{x}_{new})$ using Bayes Rule.

Solution: $p(y_{new} = i \mid \vec{x}_{new}) = \frac{p(\vec{x}_{new} | y_{new} = i) p(y_{new} = i)}{p(\vec{x}_{new})}$

(c) Using your result from part (b) write down a final mathematical equation (or pseudocode) for computing $p(y_{new} = 0 \mid \vec{x}_{new})$ and $p(y_{new} = 1 \mid \vec{x}_{new})$. **Hint:** A correct answer should involved the PDF of a Gaussian random variable, and incorporate all model parameters $\mu_{i,j}$ and $\sigma^2_{i,j}$.

Solution: We compute the following quantities for $i = 0$ and $i = 1$:

- $p(y_{new} = 1) = p$ and $p(y_{new} = 0) = 1 - p$.
- $p(\vec{x}_{new} \mid y_{new} = i) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2_{i,j}}} e^{-(\vec{x}_{new}[j] - \mu_{i,j})^2 / 2\sigma^2_{i,j}}$.
- $p(\vec{x}_{new}) = p \cdot p(\vec{x}_{new} \mid y_{new} = 1) + (1 - p) \cdot p(\vec{x}_{new} \mid y_{new} = 0)$.

Combining these three equations as in part (b) lets us compute $p(y_{new} = i \mid \vec{x}_{new})$.

(d) How can your answer from part (c) be simplified if you only seek to compute $C \cdot p(y_{new} \mid \vec{x}_{new})$ for some constant C you choose? What if you only seek to compute $B \cdot \log(C \cdot p(y_{new} \mid \vec{x}_{new}))$ for some constants B, C you choose? Can either or both of these simplified expression be used in deciding on the MAP estimate for y_{new}?

Solution: Any simplified expression of this form can be used to compare $p(y_{new} = 0 \mid \vec{x}_{new})$ and $p(y_{new} = 1 \mid \vec{x}_{new})$ since Cx and $B\log(Cx)$ are monotonic (order preserving) functions as long as C and B are positive.

So we proceed: $p(\vec{x}_{new})$ is a constant which does not depend on y_{new} and we can also factor out a constant of $\sqrt{1/2\pi}^d$. So overall we have:

$$C \cdot p(y_{new} = i \mid \vec{x}_{new}) = p(y_{new} = i) \cdot \prod_{j=1}^{d} \frac{1}{\sqrt{\sigma_{i,j}^2}} e^{-(\vec{x}_{new}[j] - \mu_{i,j})^2 / 2\sigma_{i,j}^2} \tag{2}$$

Taking logs we can further simplify:

$$\log(C \cdot p(y_{new} = i \mid \vec{x}_{new})) = \log(p(y_{new} = i)) - \sum_{j=1}^{d} \log(\sigma_{i,j}) - \sum_{j=1}^{d} (\vec{x}_{new}[j] - \mu_{i,j})^2 / 2\sigma_{i,j}^2. \tag{3}$$

This final equation is what you can use for Lab 4.