Welcome to CS109A

Gili Rusak

Agenda

- Logistic Regression*
- Bootstrapping Example*
- Naive Bayes*
- Bivariate Normal distribution*

* Review for Quiz 3

Logistic Regression Prediction

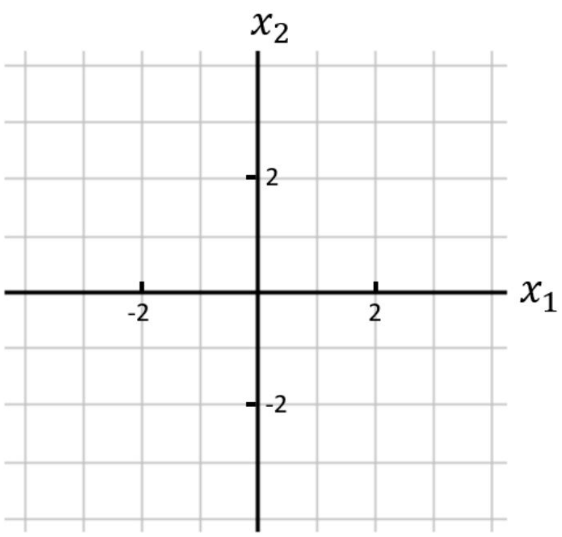
Classification is the task of choosing a value of y that maximizes $P(Y|\mathbf{X})$. Naïve Bayes worked by approximating that probability using the naïve assumption that each feature was independent given the class label.

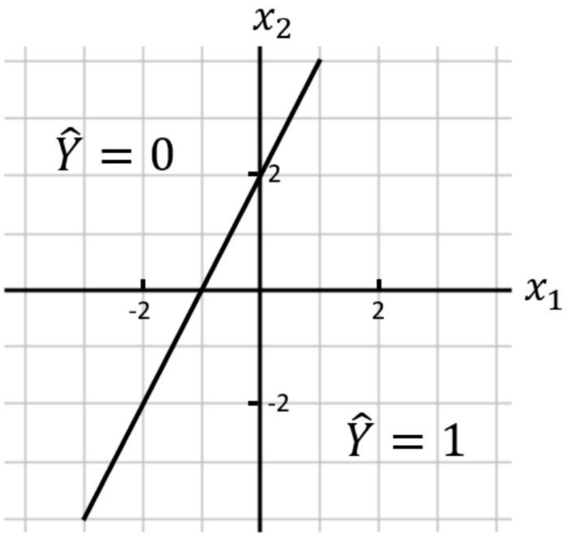
For all classification algorithms you are given n I.I.D. training datapoints $(\mathbf{x}^{(1)}, y^{(1)})$, $(\mathbf{x}^{(2)}, y^{(2)})$, $\ldots$ $(\mathbf{x}^{(n)}, y^{(n)})$ where each “feature" vector $\mathbf{x}^{(i)}$ has $m = |\mathbf{x}^{(i)}|$ features.

Logistic Regression Prediction

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \sigma(z) \text{ where } z = \theta_0 + \sum_{j=1}^{m} \theta_j x_j$$

Logistic Regression Prediction





The two parts of this problem are unrelated.

a. **Prediction.** Suppose you have trained a logistic regression classifier that accepts as input a data point (x_1, x_2) and predicts a class label $\hat{Y}$. The parameters of the model are $(\theta_0, \theta_1, \theta_2) = (2, 2, -1)$. On the axes, draw the decision boundary $\theta^T \mathbf{x} = 0$ and clearly mark which side of the boundary predicts $\hat{Y} = 0$ and which side predicts $\hat{Y} = 1$.

Logistic Regression Prediction

$\theta^T \mathbf{x}$ can be expanded as $2 + 2x_1 - x_2 = 0$ because $x_0 = 1$ by definition. The prediction is 1 when $\theta^T \mathbf{x} > 0$. For example, the origin $(x_1, x_2) = (0, 0)$ yields $\theta^T \mathbf{x} = 2$, which gives us the prediction $\hat{Y} = 1$.

See the graph above, to the right of the original.

Logistic Regression Training

$$LL(\theta) = \sum_{i=1}^{n} \log\big(f(\mathbf{x}^{(i)}, y^{(i)}|\theta)\big) = \sum_{i=1}^{n} \log\big(f(\mathbf{x}^{(i)}|\theta)P(y^{(i)}|\mathbf{x}^{(i)}, \theta)\big) \quad \text{Chain rule}$$

$$= \sum_{i=1}^{n} \log\big(f(\mathbf{x}^{(i)})f(y^{(i)}|\mathbf{x}^{(i)}, \theta)\big) \quad \mathbf{X}, \theta \text{ independent}$$

$$\theta_{MLE} = \underset{\theta}{\text{argmax}} LL(\theta) = \underset{\theta}{\text{argmax}}\left(\sum_{i=1}^{n} \log f(\mathbf{x}^{(i)}) + \log f(y^{(i)}|\mathbf{x}^{(i)}, \theta)\right) \quad \text{Log of products}$$

$$= \underset{\theta}{\text{argmax}}\left(\sum_{i=1}^{n} \log f(y^{(i)}|\mathbf{x}^{(i)}, \theta)\right) \quad \text{Constants w.r.t. } \theta$$

Logistic Regression Training

```
initialize θ_j = 0 for 0 ≤ j ≤ m
repeat many times:
    gradient[j] = 0 for 0 ≤ j ≤ m
    for each training example (x, y):
        for each 0 ≤ j ≤ m:
            gradient[j] +=  [y − 1/(1 + e^(−θ^T x))] x_j
    θ_j += η  * gradient[j] for all 0 ≤ j ≤ m
```

$$\text{gradient}[j] \mathrel{+}= \left[y - \frac{1}{1 + e^{-\theta^T \boldsymbol{x}}}\right] x_j$$

Logistic Regression Training

b. **Training.** The logistic regression parameter update equation is

$$\theta_j^{\text{new}} = \theta_j^{\text{old}} + \eta \sum_{i=1}^{n} \left[y^{(i)} - \sigma\left(\theta^{\text{old}^T} \mathbf{x}^{(i)} \right) \right] x_j^{(i)}$$

Your training set consists of two data points $\left(x_1^{(1)}, y^{(1)}\right) = (1, 1)$ and $\left(x_1^{(2)}, y^{(2)}\right) = (-1, 0)$. Given $\left(\theta_0^{\text{old}}, \theta_1^{\text{old}}\right) = (0, 0)$ and $\eta = 0.1$, find $\left(\theta_0^{\text{new}}, \theta_1^{\text{new}}\right)$.

Logistic Regression Training Solution

First notice that $\left(\theta_0^{\text{old}}, \theta_1^{\text{old}}\right) = (0, 0)$ implies that $\sigma\left({\theta^{\text{old}}}^T \mathbf{x}^{(i)}\right) = \sigma(0) = 0.5$. Therefore,

$$\begin{aligned}
\theta_0^{\text{new}} &= 0 + 0.1\left([1 - 0.5]\,(1) + [0 - 0.5]\,(1)\right) \qquad \text{since } x_0^{(i)} = 1 \text{ by definition} \\
&= 0 + 0.1(0.5 - 0.5) = 0 \\
\theta_1^{\text{new}} &= 0 + 0.1\left([1 - 0.5]\,(1) + [0 - 0.5]\,(-1)\right) \\
&= 0 + 0.1(0.5 + 0.5) = 0.1
\end{aligned}$$

Naive Bayes Example

Suppose we observe two discrete input variables X_1 and X_2 and want to predict a single binary output variable Y (which can have values 0 or 1). We know that the functional forms for the input variables are $X_1 \sim \text{Poi}(\lambda)$ and $X_2 \sim \text{Ber}(p)$, but we are not given the values of the parameters λ or p. We are, however, given a dataset of 9 training instances (shown at right.)

X_1	X_2	Y	X_1	X_2	Y
1	1	**0**	3	1	**1**
3	0	**0**	5	0	**1**
7	1	**0**	5	1	**1**
9	0	**0**	5	1	**1**
			7	1	**1**

a. Use Maximum Likelihood Estimation to estimate the parameters λ and p in the case where $Y = 0$ as well as the case $Y = 1$. You should have four parameter estimates: λ_0 and p_0 for when $Y = 0$, and λ_1 and p_1 for when $Y = 1$.

Naive Bayes Example

$$\lambda_0 = \frac{1}{4}(1 + 3 + 7 + 9) = \frac{20}{4} = 5$$

$$\lambda_1 = \frac{1}{5}(3 + 5 + 5 + 5 + 7) = \frac{25}{5} = 5$$

$$p_0 = \frac{1}{4}(1 + 0 + 1 + 0) = \frac{1}{2}$$

$$p_1 = \frac{1}{5}(1 + 0 + 1 + 1 + 1) = \frac{4}{5}$$

Naive Bayes Example

b. Use Maximum Likelihood Estimation to estimate the probability $P(Y = 1)$.

Naive Bayes Example

b. Use Maximum Likelihood Estimation to estimate the probability $P(Y = 1)$.

$P(Y = 1) = 5/9$

Naive Bayes Example

c. You observe the following testing instance: $(X_1, X_2) = (2, 0)$. Using the Naïve Bayes assumption, predict the output Y for the testing instance. For this problem, showing how you computed your prediction is worth more points than the final answer.

Naive Bayes Example

We predict $Y = 0$ if the following Naïve Bayes inequality holds:

$$P(Y=1)P(X_1=2|Y=1)P(X_2=0|Y=1) \overset{?}{<} P(Y=0)P(X_1=2|Y=0)P(X_2=0|Y=0)$$

$$\frac{5}{9}\left(\frac{\lambda_1^2}{2!}e^{-\lambda_1}\right)\left(1-\frac{4}{5}\right) \overset{?}{<} \frac{4}{9}\left(\frac{\lambda_0^2}{2!}e^{-\lambda_0}\right)\left(1-\frac{1}{2}\right)$$

$$\frac{5}{9}\left(\frac{5^2}{2!}e^{-5}\right)\frac{1}{5} \overset{?}{<} \frac{4}{9}\left(\frac{5^2}{2!}e^{-5}\right)\frac{1}{2}$$

$$\frac{5}{9}\cdot\frac{1}{5} \overset{?}{<} \frac{4}{9}\cdot\frac{1}{2}$$

$$\frac{1}{9} < \frac{2}{9}$$

Since the last inequality is true, that means the first inequality was true, so we predict $Y = 0$.

Bootstrapping Example

You are the owner of a company that makes delicious candies. The candy color Y can be red ($Y = 0$) or blue ($Y = 1$). You have two factories which produce this candy. You sample 500 candies from each factory and get the table shown at right.

Counts	Factory 1	Factory 2
$Y = 0$ (red)	260	220
$Y = 1$ (blue)	240	280

a. (6 points) What are the sample means $\bar{Y}_1$ and $\bar{Y}_2$ for the two factories?

Bootstrapping Example Solution

$$\bar{Y}_1 = \frac{260}{500}(0) + \frac{240}{500}(1) = 0.48$$
$$\bar{Y}_2 = \frac{220}{500}(0) + \frac{280}{500}(1) = 0.56$$

Bootstrapping Example

b. (7 points) Suppose you perform bootstrapping with the Factory 1 sample only. What is the probability that a bootstrap resample from Factory 1 contains at least one blue candy ($Y = 1$)? Remember that when bootstrapping you resample **with replacement** and draw a number of samples **equal to the original sample size.**

Bootstrapping Example

The probability that a single candy from the Factory 1 sample is blue, $P(Y = 1) = \frac{240}{500} = 0.48$. So the probability that there is at least one blue candy in a bootstrap resample of size 500 is

$$1 - (1 - 0.48)^{500} = 1 - 0.52^{500}$$

Bivariate Normal Distribution

Let X, Y, and Z be independent Normal variables with means of $\mu_X = 4$, $\mu_Y = 5$, and $\mu_Z = 6$ and variances $\sigma_X^2 = 16$, $\sigma_Y^2 = 25$, and $\sigma_Z^2 = 36$. If we assume $A = X + Y$ and $B = Y + Z$ are each sums of independent Normal variables, then what is the joint distribution of A and B? Restated, what is their Bivariate Normal distribution?

Bivariate Normal Distribution

$$(A, B) \sim N(\mu, \Sigma), \mu = \begin{bmatrix} \mu_X + \mu_Y \\ \mu_Y + \mu_Z \end{bmatrix}, \Sigma = \begin{bmatrix} Var(A) & Cov(A, B) \\ Cov(A, B) & Var(B) \end{bmatrix}$$

Now, $Var(A) = Var(X+Y)$, and because X and Y are independent, $Var(A) = Var(X+Y) = \sigma_X^2 + \sigma_Y^2$. Similarly, $Var(B) = Var(Y+Z) = \sigma_Y^2 + \sigma_Z^2$. Also, $Cov(A, B) = Cov(X+Y, Y+Z)$, but because X, Y, and Z are independent, $Cov(A, B) = Cov(X+Y, Y+Z) = Cov(Y, Y) = \sigma_Y^2$. Therefore,

$$\mu = \begin{bmatrix} \mu_X + \mu_Y \\ \mu_Y + \mu_Z \end{bmatrix} = \begin{bmatrix} 9 \\ 11 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_X^2 + \sigma_Y^2 & \sigma_Y^2 \\ \sigma_Y^2 & \sigma_Y^2 + \sigma_Z^2 \end{bmatrix} = \begin{bmatrix} 41 & 25 \\ 25 & 61 \end{bmatrix}$$

Thanks for a great quarter!