Naive Bayes Classifier

Sungjoo Ha

October 30th, 2013

Prediction

- Key quantity is $p(y|x)$
- Two approaches to model this
  - Discriminative
  - Generative

Generative Method

- Estimate $p(x, y)$ using D
- $p(y|x) = \frac{p(x,y)}{p(x)}$

Generative Method

Spam Filtering

- Think of a model "generating" the samples
  1. Choose spam/ham
  2. Given the decision (spam), choose the features (words)
- More formally, $p(x, y) = p(x|y)p(y)$
  1. $p(y)$ is the probability of spam/ham
  2. $p(x|y)$ is the probability of words given spam/ham

Naive Bayes

Setup

- Given $D = ((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}))$
  - $x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)}) \in \mathbb{R}^d$
  - $y^{(i)} \in \mathscr{Y}, \mathscr{Y} = \{1, \ldots, m\}$
- Assume a family of distribution p_θ s.t. for $x \in \mathbb{R}^d, y \in \mathscr{Y}$,

$$\begin{aligned} p_\theta(x, y) &= p_\theta(x|y)p_\theta(y) \\ &= p_\theta(x_1|y)p_\theta(x_2|y, x_1)p_\theta(x_3|y, x_1, x_2)\cdots \\ &\boxed{=} p_\theta(x_1|y)\cdots p_\theta(x_d|y)p_\theta(y) \end{aligned}$$

- Naive conditional independence assumption
- If $(X, Y) \sim p_\theta$, then $X_1, \ldots, X_d$ are conditionally independent given Y

Naive Bayes

Goal & Algorithm

For new $x \in \mathbb{R}^d$, predict it's y

- Estimate $\hat{\theta}$ from D
- Compute $\hat{y} \in \operatorname*{argmax}_{y \in \mathscr{Y}} p_{\hat{\theta}}(y|x)$

$$
\begin{aligned}
\hat{y} \quad &\in \quad \operatorname*{argmax}_{y \in \mathscr{Y}} p_{\hat{\theta}}(y|x) \\
&\boxed{=} \quad \operatorname*{argmax}_{y} \frac{p_{\hat{\theta}}(x|y)p_{\hat{\theta}}(y)}{p_{\hat{\theta}}(x)} \\
&= \quad \operatorname*{argmax}_{y} p_{\hat{\theta}}(x|y)p_{\hat{\theta}}(y) \\
&\boxed{=} \quad \operatorname*{argmax}_{y} p_{\hat{\theta}}(x_1|y) \cdots p_{\hat{\theta}}(x_d|y)p_{\hat{\theta}}(y)
\end{aligned}
$$

- “Bayes”
- “Naive”

Choosing p_θ

- Define the marginal distribution $p_\theta(y)$
- Define the conditional distribution $p_\theta(x|y)$

- $p_\theta(y) = p_\theta(Y = y) = \pi_y$
  - $\pi = (\pi_1, \ldots, \pi_m)$
- $p_\theta(x_i|y) = p_\theta(X_i = x_i|Y = y)$
  - If X_i is finite, e.g. $p_\theta = q(x_i, y)$
  - If X_i is countably infinite, e.g. Poisson, Geometric, etc.
  - If X_i is uncountably infinite, e.g. Gaussian, Gamma, etc.
- $\theta = (\text{all params of the distribution}, \pi)$

Estimating θ

- MLE
- MAP
- “Bayesian” – integrate out θ

Conditional Independence

$$p_\theta(x|y) = p_\theta(x_1|y) \cdots p_\theta(x_d|y)$$

- Can estimate θ more accurately with less data
  - Assuming x, y being binary
  - Joint probability requires $O(2^d)$ parameters
  - Conditional independence assumption leads to only $O(2d)$ parameters
- Wrong but simple model can work better than correct but complicated model

Spam Filter

Setup

- Given emails with label ham/spam
  - $D, (x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$
  - Label is binary decision
  - An email is described by binary word vector
    - $x =$ “Is this a spam mail?” $= [0, 1, 0, 0, 1, \dots]$
    - $p(Y = 1|X = x)$ vs. $p(Y = 0|X = x)$

Spam Filter

Algorithm

- Estimate $\hat{\theta}$ from D
  - Need to estimate

$$\hat{\theta} \quad = \quad (p(Y=1), p(X_i=1|Y=1), p(X_i=1|Y=0))$$

  - Use maximum likelihood estimate
    - For PMF on a finite set, $\theta_{MLE} = (\frac{n_1}{n}, \ldots, \frac{n_m}{n})$
  - $p(Y=1) = \frac{\#\{\text{Spams}\}}{\#\{\text{All mails}\}}$
  - $p(X_i=1|Y=1) = \frac{\#\{\text{Occurrence of word } X_i \text{ in spam}\}}{\#\{\text{Spams}\}}$
  - $p(X_i=1|Y=0) = \frac{\#\{\text{Occurrence of word } X_i \text{ in ham}\}}{\#\{\text{Hams}\}}$

Spam Filter

Goal

- For new x, predict it's y
  - $x =$ "Is this a spam mail?" $= [0, 1, 0, 0, 1, \ldots]$
  - $p(Y = 1|X = x)$ vs. $p(Y = 0|X = x)$

$$
\begin{aligned}
p(1|x) &\propto p(X_1 = 0|Y = 1)p(X_2 = 1|Y = 1) \cdots p(Y = 1) \\
p(0|x) &\propto p(X_1 = 0|Y = 0)p(X_2 = 1|Y = 0) \cdots p(Y = 0)
\end{aligned}
$$