Naive Bayesian Classifier

AW

Lecture Overview

1. Assumptions of Bayesian Classifier
2. Estimating Feature Probabilities

Bayes Predictor

- Approach so far was distribution free learning: no assumptions on the underlying distribution over the data
  - As a consequence - discriminative approach in which our goal is not to learn the underlying distribution but rather to learn an accurate predictor
- Change of strategy: generative approach, in which it is assumed that the underlying distribution over the data has a specific parametric form and our goal is to estimate the parameters of the model
  - This task is called parametric density estimation.
- If we succeed in learning the underlying distribution $\mathscr{D}$ over $X \times \{0, 1\}$ accurately, then we can predict by using the Bayes optimal classifier:

$$f_{\mathscr{D}}(x) = \begin{cases} 1 \text{ if } \Pr_{\mathscr{D}}(y = 1|x) \geq 1/2 \\ 0 \text{ otherwise} \end{cases} = \begin{cases} 1 \text{ if } \frac{\Pr_{\mathscr{D}}(y=1|x)}{\Pr_{\mathscr{D}}(y=0|x)} \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

for proof of optimality see materials section

- Another way to describe Bayes predictor for a data point $\overline{x} \in X$ is $h_{Bayes} = \arg\max_{y \in \{0,1\}} \Pr_{\mathscr{D}}(Y = y|X = \overline{x})$

Conditional Independence of Features

- Let X have n features with respective finite domains $\mathbb{D}_1, \ldots, \mathbb{D}_n$
- $$\begin{aligned} \Pr_{\mathscr{D}}(Y = y | X = \overline{x}) &= \Pr_{\mathscr{D}}(Y = y | x_1 = a_1, \ldots, x_n = a_n) \\ &= \frac{\Pr_{\mathscr{D}}(x_1 = a_1, \ldots, x_n = a_n | Y = y) \Pr_{\mathscr{D}}(Y = y)}{\Pr_{\mathscr{D}}(x_1 = a_1, \ldots, x_n = a_n)} \end{aligned}.$$
- Bayes predictor $h_{Bayes} = \arg\max_{y \in \{0,1\}} \Pr_{\mathscr{D}}(Y = y | X = \overline{x})$ would need to be optimized with respect to $d = |\mathbb{D}_1| \times \cdots \times |\mathbb{D}_n|$ parameters $\Pr_{\mathscr{D}}(x_1 = a_1, \ldots, x_n = a_n | Y = y)$
  - For example if all features are Boolean then $d = 2^n$. So the number of parameters grows exponentially with the number of features - a bit too much!
- The Naive Bayes approach makes the assumption about distribution (aka *generative assumption*) that given a class, all features are independent of each other given class y, i.e. $\Pr_{\mathscr{D}}(x_1 = a_1, \ldots, x_n = a_n | Y = y) = \prod_{i=1}^{n} \Pr(x_i = a_i | Y = y)$
- For 2 class predictor under these assumptions we have
$$f_{\mathscr{D}}(x) = \begin{cases} 1 \text{ if } \frac{\Pr_{\mathscr{D}}(y=1|x)}{\Pr_{\mathscr{D}}(y=0|x)} = \frac{\Pr(Y=1) \prod_{i=1}^{n} \Pr(x_i = a_i | Y=1)}{\Pr(Y=0) \prod_{i=1}^{n} \Pr(x_i = a_i | Y=0)} \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

Lecture Overview

1. Assumptions of Bayesian Classifier
2. Estimating Feature Probabilities

Likelihood of a Parameter

We need to estimate $\Pr(x_i = a_i | Y = y)$ and $\Pr(Y = y)$. From our training data for feature x_i we have a sample S_i^0 of pairs $(s_{i_1}, 0), \ldots, (s_{i_k}, 0)$ of class 0 and a sample S_i^1 of pairs $(s_{i_{k+1}}, 1), \ldots (s_{i_m}, 1)$ of class 1. We make more assumptions:

- Our samples are coming from unknown pdf (or pmf) $f_0(\cdot) = \mathscr{D}|_{X_i \times Y}$ (where $\mathscr{D} \sim X \times Y$) that belongs to a certain family of distributions $\{f(\cdot|\theta), \theta \in \Theta\}$, so $f_0(\cdot)$ is in fact $f_0(\cdot|\theta_0)$ for some parameter θ_0 that we need to estimate
- $S_i^0 = \{s_{i_1}, \ldots, s_{i_k}\}$ (resp. S_i^1) is drawn independently from $\mathscr{D}|_{X,0}$

Let $f(s_{i_1}, s_{i_2}, \ldots, s_{i_k} \mid \theta) = f(s_{i_1} \mid \theta) \times f(s_{i_2} \mid \theta) \times \cdots \times f(s_{i_k} \mid \theta)$. In this joint pdf (pmf)

- Observed values $s_{i_1}, \ldots, s_{i_k}$ are considered fixed ”parameters”
- θ is the function’s free variable

Define log likelihood (or when more convenient $\ln$-likelihood)

$$\begin{aligned}
\mathscr{L}(\theta | s_{i_1}, \ldots, s_{i_k}) &= \log\left(f(s_{i_1}, s_{i_2}, \ldots, s_{i_n} \mid \theta)\right) \\
&= \log\left(\prod_{j=1}^{k} f(s_{i_j} \mid \theta)\right) = \sum_{j=1}^{k} \log(f(s_{i_j} \mid \theta))
\end{aligned}$$

Maximum Likelihood Estimation

Given a sample S maximum likelihood estimator of parameter θ is

$$\hat{\theta} = \arg\max_{\theta} \mathscr{L}(\theta|S)$$

For our purposes:

Maximum Likelihood Estimation

Given a sample S maximum likelihood estimator of parameter θ is

$$\hat{\theta} = \arg\max_\theta \mathscr{L}(\theta|S)$$

For our purposes:

Estimate $\Pr(Y = y)$.

- We assume Bernoulli distribution with unknown probability θ of 1. Given sample $S = \{y_1, \ldots, y_m\}$
- Then $f(y_i|\theta) = \begin{cases} 1 - \theta \text{ if } y_i = 0 \\ \theta \text{ otherwise} \end{cases}$
- $$\begin{aligned} \mathscr{L}(\theta|S) &= \log\left(\prod_{i=1}^m \theta^{y_i} \cdot (1-\theta)^{1-y_i}\right) \\ &= \log(\theta)\sum_{i=1}^m y_i + \log(1-\theta)\sum_{i=1}^m (1-y_i) \end{aligned}$$
- $\frac{d}{d\theta}(\mathscr{L}(\theta|S)) = \frac{\sum_{i=1}^m y_i}{\theta \ln 2} - \frac{\sum_{i=1}^m (1-y_i)}{(1-\theta)\ln 2} = 0$ so $\theta = \frac{1}{m}\sum_{i=1}^m y_i$

Maximum Likelihood Estimation

Given a sample S maximum likelihood estimator of parameter θ is

$$\hat{\theta} = \arg\max_{\theta} \mathscr{L}(\theta|S)$$

For our purposes:

Estimate $\Pr(X = x \mid Y = 0)$ (or resp. $\Pr(X = x \mid Y = 1)$) when feature X has finite domain $[d] = \{0, 1, \ldots, d\}$ and the sample $S = S_0 \cup S_1$ where $S_0 = \{(z_1, 0) \ldots, (z_k, 0)\}$ and $S_1 = \{(z_{k+1}, 1) \ldots, (z_m, 1)\}$.

- Let our subsample S_0 consists of x_0 occurrences of $X = 0$, x_1 occurrences of $X = 1$, $\ldots$, x_d occurrences of $X = d$
- By law of total probability $\Pr(Y = 0) = \sum_{i=0}^{d} \Pr(X = i \wedge Y = 0) = \sum_{i=0}^{d} \Pr(X = i \mid Y = 0) \Pr(Y = 0)$ so $\sum_{i=0}^{d} \Pr(X = i \mid Y = 0) = 1$. Let $\theta_i = \Pr(X = i \mid Y = 0)$. So $\sum_i \theta_i = 1$.
- Then $f(S \mid \theta_1, \ldots, \theta_d) = f(\{z_1, \ldots, z_k\}) \mid \theta_1, \ldots, \theta_d) = \prod_{i=0}^{d} (\theta_i \cdot \Pr(Y = 0))^{x_i}$
- The likelihood of our sample S then is

$$\mathscr{L}(\theta|S) \quad = \quad \log \left(\prod_{i=0}^{d} (\theta_i \cdot \Pr(Y = 0))^{x_i} \right)$$

Maximum Likelihood Estimation

Given a sample S maximum likelihood estimator of parameter θ is

$$\hat{\theta} = \arg\max_{\theta} \mathscr{L}(\theta|S)$$

For our purposes:
Estimate $\Pr(X = x \mid Y = 0)$ (or resp. $\Pr(X = x \mid Y = 1)$) when feature X has finite domain $[d] = \{0, 1, \ldots, d\}$ and the sample $S = S_0 \cup S_1$ where $S_0 = \{(z_1, 0) \ldots, (z_k, 0)\}$ and $S_1 = \{(z_{k+1}, 1) \ldots, (z_m, 1)\}$.

- Sample S_0 consists of x_i occurrences of $X = i$. Denote $\theta_i = \Pr(X = i \mid Y = 0)$. So $\sum_i \theta_i = 1$.
- The likelihood of our sample S then is

$$\begin{aligned} \mathscr{L}(\theta|S) &= \log\left(\prod_{i=0}^{d} (\theta_i \cdot \Pr(Y = 0))^{x_i}\right) \\ &= k \log(\Pr(Y = 0)) + \sum_{i=0}^{d} x_i \log(\theta_i) \end{aligned}$$

- Thus maximize $\mathscr{L}(\theta|S)$ is to maximize $\sum_{i=0}^{d} x_i \log(\theta_i)$ since first term doesn't depend on parameters. Equivalently this means to minimize $-\sum_{i=0}^{d} x_i \log(\theta_i)$.

Maximum Likelihood Estimation

Given a sample S maximum likelihood estimator of parameter θ is

$$\hat{\theta} = \arg\max_\theta \mathscr{L}(\theta|S)$$

For our purposes:
Estimate $\Pr(X = x \mid Y = 0)$ (or resp. $\Pr(X = x \mid Y = 1)$) when feature X has finite domain $[d] = \{0, 1, \dots, d\}$ and the sample $S = S_0 \cup S_1$ where $S_0 = \{(z_1, 0) \dots, (z_k, 0)\}$ and $S_1 = \{(z_{k+1}, 1) \dots, (z_m, 1)\}$. Let sample S_0 consists of x_i occurrences of $X = i$ and $\theta_i = \Pr(X = i \mid Y = 0)$. So $\sum_i \theta_i = 1$.
To maximize $\mathscr{L}(\theta|S)$ we must solve the optimization problem:

$$\begin{array}{ll} \text{minimize} & -\sum_{i=0}^{d} x_i \log(\theta_i) \\ \text{subject to} & \\ & \sum_{i=0}^{d} \theta_i - 1 = 0 \end{array}$$

We use Lagrangian method: introduce a new variables λ_i, one per constraint, called a Lagrange multipliers and study a Lagrangian function $\mathcal{L}(x_1, \dots x_n, \lambda_1, \dots, \lambda_k) = f(x_1, \dots, x_n) \pm \sum_{i=1}^{k} \lambda_i \cdot g_i(x_1, \dots, x_n)$ where $f(x_1, \dots, x_n)$ is the goal and $g_i(x_1, \dots, x_n)$ are constraint functions

Maximum Likelihood Estimation

Given a sample S maximum likelihood estimator of parameter θ is

$$\hat{\theta} = \arg\max_\theta \mathscr{L}(\theta|S)$$

For our purposes:

$\boxed{\text{Estimate } \Pr(X = x \mid Y = 0)}$ (or resp. $\Pr(X = x \mid Y = 1)$) when feature X has finite domain $[d] = \{0, 1, \ldots, d\}$ and the sample $S = S_0 \cup S_1$ where $S_0 = \{(z_1, 0) \ldots, (z_k, 0)\}$ and $S_1 = \{(z_{k+1}, 1) \ldots, (z_m, 1)\}$. Let sample S_0 consists of x_i occurrences of $X = i$ and $\theta_i = \Pr(X = i \mid Y = 0)$. So $\sum_i \theta_i = 1$.

To maximize $\mathscr{L}(\theta|S)$ we must solve the optimization problem:

$$\begin{array}{ll} \text{minimize} & -\sum_{i=0}^{d} x_i \log(\theta_i) \\ \text{subject to} & \\ & \sum_{i=0}^{d} \theta_i - 1 = 0 \end{array}$$

In our case $\mathcal{L}(\theta_0, \ldots, \theta_d, \lambda) = -\sum_{i=0}^{d} x_i \log(\theta_i) + \lambda\left(\sum_{i=0}^{d} \theta_i - 1\right)$

Stationary point of $\mathcal{L}$ is a point where the partial derivatives of $\mathcal{L}$ are zero. If $f(x_1^*, \ldots, x_n^*)$ is a minimum (resp. maximum) of the goal for the original constrained problem, then there exists $\lambda_1^*, \ldots, \lambda_k^*$ such that $(x_1^*, \ldots, x_n^*, \lambda_1^*, \ldots, \lambda_k^*)$ is a stationary point for the Lagrangian.

Maximum Likelihood Estimation

Given a sample S maximum likelihood estimator of parameter θ is

$$\hat{\theta} = \arg\max_{\theta} \mathscr{L}(\theta|S)$$

For our purposes:

Estimate $\Pr(X = x \mid Y = 0)$ (or resp. $\Pr(X = x \mid Y = 1)$) when feature X has finite domain $[d] = \{0, 1, \ldots, d\}$ and the sample $S = S_0 \cup S_1$ where $S_0 = \{(z_1, 0) \ldots, (z_k, 0)\}$ and $S_1 = \{(z_{k+1}, 1) \ldots, (z_m, 1)\}$. Let sample S_0 consists of x_i occurrences of $X = i$ and $\theta_i = \Pr(X = i \mid Y = 0)$. So $\sum_i \theta_i = 1$.

To maximize $\mathscr{L}(\theta|S)$ we must solve the optimization problem:

$$\begin{array}{ll} \text{minimize} & -\sum_{i=0}^{d} x_i \log(\theta_i) \\ \text{subject to} & \\ & \sum_{i=0}^{d} \theta_i - 1 = 0 \end{array}$$

So $\max_{\lambda} \min_{\theta_1,\ldots,\theta_d} \left[-\sum_{i=0}^{d} x_i \log(\theta_i) + \lambda \left(\sum_{i=0}^{d} \theta_i - 1 \right) \right]$. Partial derivatives for θ_i is $\frac{x_i}{\theta_i} - \lambda = 0$ for all i, or $\theta_i = \frac{x_i}{\lambda}$. Partial derivative for λ is $\sum_{i=0}^{d} \theta_i - 1 = 0$ which yields $\lambda = \sum_{i=0}^{d} x_i = \begin{cases} k \text{ if } Y = 0 \\ m - k \text{ if } Y = 1 \end{cases}$, so e.g. for $Y = 0$ we get $\theta_i = \frac{x_i}{k}$.

Maximum Likelihood Estimation

Given a sample S maximum likelihood estimator of parameter θ is

$$\hat{\theta} = \arg\max_{\theta} \mathscr{L}(\theta|S)$$

For our purposes:
Estimate $\Pr(X = x \mid Y = 0)$ (or resp. $\Pr(X = x \mid Y = 1)$) when feature X has continuous domain, assume that $\Pr(X = x \mid Y = 0)$ is normally distributed and the sample is $S_0 = \{(z_1, 0) \ldots, (z_k, 0)\}$. In other words we know pdf $p(X = x \mid Y = 0) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ where parameter $\overline{\theta}$ is $(\mu, \sigma)^T$. Then $f(\{z_1, \ldots, z_k\} \mid \overline{\theta}) = \prod_{i=1}^{k} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right)$ and

$$\begin{aligned}
\mathscr{L}(\overline{\theta}|S) &= \log\left(\prod_{i=1}^{k} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right)\right) \\
&= -\frac{\sum_{i=1}^{k}(z_i-\mu)^2}{2\sigma^2} - k\log\left(\sigma\sqrt{2\pi}\right)
\end{aligned}$$

So obtaining $\hat{\theta}$ involves solving $\frac{\partial}{\partial\mu}\left(\mathscr{L}(\overline{\theta}|S)\right) = 0$ and $\frac{\partial}{\partial\sigma}\left(\mathscr{L}(\overline{\theta}|S)\right) = 0$,

Maximum Likelihood Estimation

Given a sample S maximum likelihood estimator of parameter θ is

$$\hat{\theta} = \arg\max_\theta \mathscr{L}(\theta|S)$$

For our purposes:

Estimate $\Pr(X = x \mid Y = 0)$ (or resp. $\Pr(X = x \mid Y = 1)$) when feature X has continuous domain, assume that $\Pr(X = x \mid Y = 0)$ is normally distributed and the sample is $S_0 = \{(z_1, 0) \ldots, (z_k, 0)\}$.

$$\begin{aligned}
\mathscr{L}(\bar{\theta}|S) &= \log\left(\prod_{i=1}^k \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right)\right) \\
&= -\frac{\sum_{i=1}^k (z_i-\mu)^2}{2\sigma^2} - k\log\left(\sigma\sqrt{2\pi}\right)
\end{aligned}$$

So obtaining $\hat{\theta}$ involves solving $\frac{\partial}{\partial\mu}\left(\mathscr{L}(\bar{\theta}|S)\right) = 0$ and $\frac{\partial}{\partial\sigma}\left(\mathscr{L}(\bar{\theta}|S)\right) = 0$,

$$\begin{aligned}
\frac{\partial}{\partial\mu}\left(\mathscr{L}(\bar{\theta}|S)\right) &= \frac{\sum_{i=1}^k (z_i-\mu)}{\sigma^2} = 0 \\
&\Downarrow \\
\hat{\mu} &= \frac{\sum_{i=1}^k z_i}{k}
\end{aligned}
\qquad
\begin{aligned}
\frac{\partial}{\partial\sigma}\left(\mathscr{L}(\bar{\theta}|S)\right) &= \frac{\sum_{i=1}^k (z_i-\mu)^2}{\sigma^3} - \frac{k}{\sigma} = 0 \\
&\Downarrow \\
\hat{\sigma} &= \sqrt{\frac{\sum_{i=1}^k (z_i-\mu)^2}{k}}
\end{aligned}$$