DM825

Introduction to Machine Learning

Lecture 7
Gaussian discriminant analysis
Naive Bayes

Marco Chiarandini

Department of Mathematics & Computer Science
University of Southern Denmark

Outline

1. Gaussian Discriminant Analysis

2. Naive Bayes
   Multi-variate Bernoulli Event Model
   Multinomial Event Model

3. Support Vector Machines

- Discriminative approach learns $p(y|x)$
- Generative approach learns $p(x|y)$

Generative Method

1. Model $p(y)$ and $p(x \mid y)$
2. learn parameters of the models by maximizing joint likelihood $p(x, y) = p(x|y)p(y)$
3. express

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)} = \frac{p(x \mid y)p(y)}{\sum_{y \in Y} p(x \mid y)p(y)}$$

4. predict

$$\begin{aligned}
\arg\max_y p(y \mid x) &= \arg\max_y \frac{p(x \mid y)p(y)}{p(x)} \\
&= \arg\max_y \frac{p(x \mid y)p(y)}{\sum_{y \in Y} p(x \mid y)p(y)} \\
&= \arg\max_y p(x \mid y)p(y)
\end{aligned}$$

Outline

1. Gaussian Discriminant Analysis

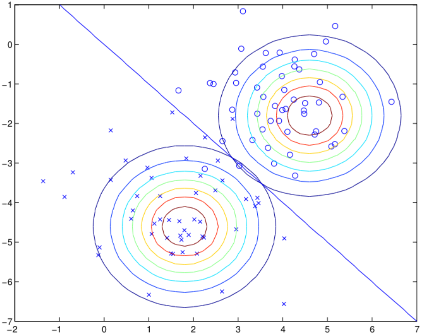
2. Naive Bayes
   - Multi-variate Bernoulli Event Model
   - Multinomial Event Model

3. Support Vector Machines

Gaussian Discriminant Analysis

Let $\vec{x}$ be a vector of continuous variables
We will assume $p(\vec{x} \mid y)$ is a multivariate Gaussian distribution

$$p(\vec{x}, \vec{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(\frac{1}{2}(\vec{x} - \vec{\mu})^T \mathbf{\Sigma}^{-1}(\vec{x} - \vec{\mu})\right)$$



Gaussian Discriminant Analysis

Step 1: we model the probabilities:

$$\begin{aligned} y &\sim \text{Bernoulli}(\varphi) \\ x|y=0 &\sim N(\mu_0, \Sigma) \\ x|y=1 &\sim N(\mu_1, \Sigma) \end{aligned}$$

that is: $p(y) = \phi^y(1-\phi)^{1-y}$
$p(x \mid y=0) = \mathcal{N}(\vec{\mu}_0, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(\frac{1}{2}(\vec{x}-\vec{\mu}_0)^T\mathbf{\Sigma}^{-1}(\vec{x}-\vec{\mu}_0)\right)$
$p(x \mid y=1) = \mathcal{N}(\vec{\mu}_1, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T\mathbf{\Sigma}^{-1}(\vec{x}-\vec{\mu}_1)\right)$
Step 2: we express the joint likelihood of a set of data $i = 1 \dots m$:

$$\begin{aligned} l(\phi, \mu_0, \mu_1, \mathbf{\Sigma}) &= \prod_{i=1}^{m} p(x^i, y^i) \\ &= \prod_{i=1}^{m} p(x^i \mid y^i)p(y^i) \end{aligned}$$

We substitute the model assumptions above and maximize $\log l(\phi, \mu_0, \mu_1, \mathbf{\Sigma})$ in $\phi, \mu_0, \mu_1, \mathbf{\Sigma}$

Solutions:

$$\phi = \frac{\sum_{i=1}^m y^i}{m} = \frac{\sum_i I\{y^i = 1\}}{m}$$
$$\mu_0 = \frac{\sum_i I\{y^i = 0\}x^i}{\sum_i I\{y^i = 0\}}$$
$$\mu_1 = \frac{\sum_i I\{y^i = 1\}x^i}{\sum_i I\{y^i = 1\}}$$
$$\boldsymbol{\Sigma} = \ldots$$

Compare with logistic regression where we maximized the conditional likelihood instead!

Step 3 and 4:

$$\arg\max_y p(y \mid x) = \arg\max_y \frac{p(x \mid y)p(y)}{p(x)}$$
$$\arg\max_y p(x \mid y)p(y)$$

Comments

- In GDA: $x \mid y \sim \text{Gaussian} \Longrightarrow$logistic posterior for $p(y=1 \mid x)$ (see sec 4.2 of B1)
- In logistic regression we model $p(y \mid x)$ as logistic also other distributions, eg:

$$\begin{aligned} x|y=0 &\sim \text{Poisson}(\lambda_0) \\ x|y=1 &\sim \text{Poisson}(\lambda_1) \end{aligned}$$

$$\begin{aligned} x|y=0 &\sim \text{ExpFam}(\eta_0) \\ x|y=1 &\sim \text{ExpFam}(\eta_1) \end{aligned}$$

- hence we make stronger assumptions in GDA. If we do not know where the data come from the logistic regression analysis would be more robust. If we know then GDA may perform better.

- When Σ is the same for all class conditional densities then the decision boundaries are linear $\leadsto$ Linear discriminative analysis (LDA)
- When the class conditional densities do not share Σ then quadratic discriminant

Outline

1. Gaussian Discriminant Analysis

2. Naive Bayes
   - Multi-variate Bernoulli Event Model
   - Multinomial Event Model

3. Support Vector Machines

Chain Rule of Probability

permits the calculation of the joint distribution of a set of random variables using only conditional probabilities.

Consider the set of events $A_1, A_2, \dots A_n$. To find the value of the joint distribution, we can apply the definition of conditional probability to obtain:

$$\Pr(A_n, A_{n-1}, \dots, A_1) = \Pr(A_n \mid A_{n-1}, \dots, A_1) \Pr(A_{n-1}, A_{n-2}, \dots, A_1)$$

repeating the process with each final term:

$$\Pr(\cap_{k=1}^{n} A_k) = \prod_{k=1}^{n} \Pr(A_k \mid \cap_{j=1}^{k} A_j)$$

For example:

$$\Pr(A_4, A_3, A_2, A_1) = \Pr(A_4|A_3, A_2, A_1) \Pr(A_3|A_2, A_1) \Pr(A_2|A_1) \Pr(A_1)$$

Multi-variate Bernoulli Event Model

We want to decide whether an email is spam $y \in \{0, 1\}$ given some discrete features $\vec{x}$.
How to represent emails by a set of features?
Binary array, each element corresponds to a word in the vocabulary and the bit indicates whether the word is present or not in the data.

$$\vec{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

$\vec{x} \in \{0, \}^n$
$n = 50000$ (large number)
2^{50000} possible bit vectors
$2^{50000-1}$ parameters to learn

We collect examples, look at those that are spam $y = 1$ and learn $p(x \mid y = 1)$, then at those $y = 0$ and learn $p(x \mid y = 0)$

Step 1: For a given example i we treat each x_j^i independently

$$\begin{aligned} p(y) &\sim \phi_y \\ \forall j : p(x_j|y=0) &\sim \phi_{j|y=0} \\ \forall j : p(x_j|y=1) &\sim \phi_{j|y=1} \end{aligned}$$

Step 2: Maximize joint likelihood
Assume x_js are conditionally independent given y. By chain rule:

$$\begin{aligned} p(x_1, \ldots, x_{50000}) &= p(x_1 \mid y) p(x_2 \mid y, x_1) p(x_3 \mid y, x_1, x_2) \ldots \\ &= p(x_1 \mid y) p(x_2 \mid y) p(x_3 \mid y) \ldots && \text{cond. indep.} \\ &= \prod_{i=1}^{m} p(x_i \mid y) \end{aligned}$$

$$
\begin{aligned}
l(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}) &= \prod_{i=1}^{m} p(\vec{x}^i, y^i) \\
&= \prod_{i=1}^{m} \prod_{j=1}^{n} p(x_j^i \mid y^i) p(y^i)
\end{aligned}
$$

Solution:

$$
\begin{aligned}
\phi_y &= \frac{\sum_i I\{y^i = 1\}}{m} \\
\phi_{j|y=1} &= \frac{\sum_i I\{y^i = 1, x_j^i = 1\}}{\sum_i I\{y^i = 1\}} \\
\phi_{j|y=0} &= \frac{\sum_i I\{y^i = 0, x_j^i = 1\}}{\sum_i I\{y^i = 0\}}
\end{aligned}
$$

Step 3 and 4: prediction as usual but remember to use logarithms

Laplace Smoothing

what if $p(x_{300000}|y=1)=0$ and $p(x_{300000}|y=0)=0$ because we do not have any observation in the training set with that word?

$$
\begin{aligned}
p(y=1|x) &= \frac{p(x \mid y=1)p(y=1)}{p(x \mid y=1)p(y=1)+p(x \mid y=0)p(y=0)} \\
&= \frac{\prod_{j=1}^{50000} p(x_j \mid y=1)p(y=1)}{\prod_{j=1}^{50000} p(x_j \mid y=1)p(y=1)+\prod_{j=1}^{50000} p(x_j \mid y=0)p(y=0)} \\
&= \frac{0}{0+0}
\end{aligned}
$$

Laplace smoothing: assume some observations

$$p(x|y)=\frac{c(x,y)+k}{c(y)+k|x|}$$

$$
\begin{aligned}
\phi_y &= \frac{\sum_i I\{y^i=1\}+1}{m+K} \\
\phi_{j|y=1} &= \frac{\sum_i I\{y^i=1, x_j^i=1\}+1}{\sum_i I\{y^i=1\}+2} \\
\phi_{j|y=0} &= \frac{\sum_i I\{y^i=0, x_j^i=1\}+1}{\sum_i I\{y^i=0\}+2}
\end{aligned}
$$

Multinomial Event Model

We look at a generalization of the previous Naive Bayes that allows to take into account also of the number of times a word appear as well as the position.

Let $x^i \in \{1, 2, \ldots K\}$, for example, a continuous variable discretized in buckets

Let $[x^i_1, x^i_2, \ldots x^i_{n_i}]$, $x^i_j \in \{1, 2, \ldots, K\}$ represent the word in position j, n_i # of word in the ith email

Step 1:

$$\begin{aligned} p(y) &\sim \phi_y \\ \forall j : p(x_j = k|y = 0) &\sim \phi_{j|y=0} \\ \forall j : p(x_j = k|y = 1) &\sim \phi_{j|y=1} \end{aligned}$$

assumed that $p(x_j = k|y = 0)$ is the same for all j

$\phi_{j|y=0}$ are parameters of multinomial Bernoulli distributions

Step 2: Joint likelihood:

$$\begin{aligned}\mathcal{L}(\phi, \phi_{k|y=0}, \phi_{k|y=1}) &= \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) \\ &= \prod_{i=1}^{m} \left(\prod_{j=1}^{n_i} p(x_j^{(i)} | y; \phi_{k|y=0}, \phi_{k|y=1}) \right) p(y^{(i)}; \phi_y).\end{aligned}$$

Solution:

$$\begin{aligned}\phi_{k|y=1} &= \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} n_i} \\ \phi_{k|y=0} &= \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} n_i} \\ \phi_y &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}.\end{aligned}$$

Laplace Smoothing

$$
\begin{aligned}
\phi_{k|y=1} &= \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} n_i + |V|} \\
\phi_{k|y=0} &= \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 0\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} n_i + |V|}.
\end{aligned}
$$

Outline

1. Gaussian Discriminant Analysis

2. Naive Bayes
    - Multi-variate Bernoulli Event Model
    - Multinomial Event Model

3. Support Vector Machines

Support Vector Machines (Intro)

Support vector machines: discriminative approach that implements a non linear decision with basis in linear classifiers.

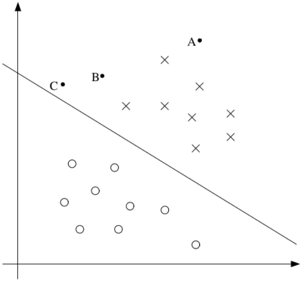
- lift points to a space where they are linearly separable
- find linear separator

Let's focus first on how to find a linear separator. Desiderata:

predict "1" iff $\vec{\theta}^T \vec{x} \geq 0$
predict "0" iff $\vec{\theta}^T \vec{x} < 0$

also wanted:

if $\vec{\theta}^T \vec{x} \gg 0$ very confident that $y = 1$
if $\vec{\theta}^T \vec{x} \ll 0$ very confident that $y = 0$



Hence it would be nice if:

$\forall i : y^i = 1$ we have $\vec{\theta}^T \vec{x} \gg 0$
$\forall i : y^i = 0$ we have $\vec{\theta}^T \vec{x} \ll 0$

Notation

Assume training set is linearly separable

Let's change notation:
$y \in \{-1, 1\}$ (instead of $\{0, 1\}$ like in GLM)
Let's have h output values $\{-1, 1\}$:

$$f(z) = \text{sign}(z) \begin{cases} 1 & if z \geq 0 \\ -1 & if z < 0 \end{cases}$$

(hence no probabilities like in logistic regression)

$h(\vec{\theta}, \vec{x}) = f(\vec{\theta}\vec{x}), \vec{x} \in \mathbb{R}^{n+1}, \vec{\theta} \in \mathbb{R}^{n+1}$
$h(\vec{\theta}, \vec{x}) = f(\vec{\theta}\vec{x} + \theta_0), \vec{x} \in \mathbb{R}^{n}, \vec{\theta} \in \mathbb{R}^{n}, \theta_0 \in \mathbb{R}$

Functional Marginal

Def.: The functional margin of a hyperplane $(\vec{\theta}, \theta_0)$ w.r.t. a specific example (x^i, y^i) is:

$$\hat{\gamma}^i = y^i(\vec{\theta}^T \vec{x}^i + \theta_0)$$

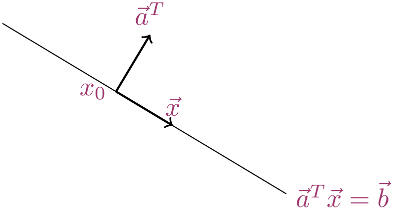
For the decision boundary $\vec{\theta}^T \vec{x} + \theta_0$ that defines the linear boundary:
we want $\vec{\theta}^T \vec{x} \gg 0$ if $y^i = +1$
we want $\vec{\theta}^T \vec{x} \ll 0$ if $y^i = -1$
If $y^i(\vec{\theta}^T \vec{x}^i + \theta_0) > 0$ then i is classified correctly.
Hence, we want to maximize $y^i(\vec{\theta}^T \vec{x}^i + \theta_0)$.
This can be achieved by maximizing the worst case for the training set

$$\boxed{\hat{\gamma} = \min_i \hat{\gamma}^i}$$

Note: scaling $\vec{\theta} \to 2\vec{\theta}, \theta_0 \to \theta_o$ would make $\hat{\gamma}$ arbitrarily large. Hence we impose: $\| \vec{\theta} \| = 1$

Hyperplanes

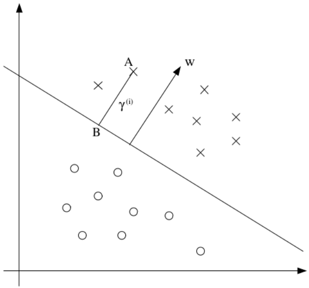
hyperplane: set of the form $\{\vec{x} \mid \vec{a}^T\vec{x} = \vec{b}\}$ ($\vec{a} \neq 0$)



- $\vec{a}$ is the normal vector
- hyperplanes are affine and convex sets

Geometric Marginal

Def. the geometric margin γ^i is the distance of example i from the linear separator, A-B



$\frac{\vec{\theta}}{\|\theta\|}$ unit vector orthogonal to the separating hyperplane

A-B: $x^i - \gamma^i \frac{\vec{\theta}}{\|\vec{\theta}\|}$

since B is on the linear separator, substituting the part above in $\vec{\theta}^T \vec{x} + \theta_0 = 0$:

$$\vec{\theta}^T \left(x^i - \gamma^i \frac{\vec{\theta}}{\| \vec{\theta} \|} \right) + \theta_0 = 0$$

Solving for γ^i:

$$\vec{\theta}^T x^i + \theta_0 = \gamma^i \frac{\vec{\theta}^T \vec{\theta}}{\| \vec{\theta} \|}$$

$$\gamma^i = \frac{\vec{\theta}}{\| \vec{\theta} \|}^T x^i + \frac{\theta_0}{\| \vec{\theta} \|}$$

This was for the positives. We can develop the same for the negatives but we would have a negative sign. Hence the quantity:

$$\gamma^i = y^i \left(\frac{\vec{\theta}}{\| \vec{\theta} \|}^T x^i + \frac{\theta_0}{\| \vec{\theta} \|} \right)$$

will be always positive.

To maximize the distance of the line from all points we maximize the worst case, that is:

$$\boxed{\gamma = \min_i \gamma^i}$$

- $\gamma = \frac{\hat{\gamma}}{\|\vec{\theta}\|}$

- Note that if $\| \vec{\theta} \| = 1$ then $\hat{\gamma}^i = \gamma^i$ the two marginal correspond

- geometric margin is invariant to scaling $\vec{\theta} \rightarrow 2\vec{\theta}, \theta_0 \rightarrow \theta_o$

Optimal margin classifier

$$\max_{\gamma\vec{\theta},\theta_0} \quad \gamma \tag{1}$$

$$\gamma \leq y^i(\vec{\theta}^T\vec{x} + \theta_0) \qquad \forall i = 1, \ldots, m \tag{2}$$

$$\parallel \vec{\theta} \parallel = 1 \tag{3}$$

(2) implements $\gamma = \min \gamma^i$
(3) is a nonconvex constraint thanks to which the two marginals are the same.