CSCI 590: Machine Learning

Lecture 5:
Bayes Rule, Naïve Bayes, Bayesian Inference, Gaussian Partitions
Instructor: Murat Dundar

Acknowledgement:
Some of these slides are taken from course textbook website http://research.microsoft.com/~cmbishop/prml/ and from the book Data Mining: Concepts and Techniques by Jiawei Han et al.
Bayes’ Theorem: Basics

- Let X be a data sample ("evidence"): class label is unknown
- Let H be a hypothesis that X belongs to class C
- Classification is to determine $P(H|X)$, the probability that the hypothesis holds given the observed data sample X
- $P(H)$ (prior probability), the initial probability
- $P(X)$: probability that sample data is observed
- $P(H|X)$ (posterior probability), the probability of observing the sample X, given that the hypothesis holds
Bayes Theorem

Given $X$, posteriori probability of a hypothesis $H$, $P(H \mid x)$, follows the Bayes theorem

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(x)}$$

Informally, this can be written as

posteriori = likelihood x prior/evidence

Predicts $X$ belongs to $C_j$ iff the probability $P(C_j \mid X)$ is the highest among all the $P(C_i \mid X)$ for all the classes

Practical difficulty: require initial knowledge of many probabilities, significant computational cost
Naïve Bayes Classifier (1)

Let \( D \) be the training set where each sample is represented by an \( d \)-dimensional attribute vector \( X = (x_1, x_2, ..., x_d) \).

Suppose there are \( m \) classes \( C_1, C_2, ..., C_m \).

Classification is to derive the maximum posteriori, i.e., the maximal \( P(C_i | X) \).

This can be derived from Bayes’ theorem

\[
P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(x)}
\]

Since \( P(X) \) is constant for all classes, only needs to be maximized

\[
P(C_i | X) = P(X | C_i)P(C_i)
\]
Naïve Bayes Classifier (2)

A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

\[
P(X \mid C_i) = \prod_{k=1}^{d} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_d \mid C_i)
\]

If attributes are categorical, \( P(x_k \mid C_i) \) is the # of tuples in \( C_i \) having value \( x_k \) for that attribute divided by \( n_i \) (# of samples of \( C_i \) in \( D \))

If attributes are continuous-valued, model \( P(x_k \mid C_i) \) by a distribution, i.e., a univariate Gaussian distribution
Naïve Bayes Classifier (3)

Advantages

- Easy to implement
- Good results obtained in some cases

Disadvantages

- Assumption: class conditional independence, therefore loss of information
  - Not a very practical assumption. Attributes are correlated in most real-world settings.

We should model $P(X|C_i)$ directly without any independence assumption.
Bayesian Inference

- Prior
- Likelihood function
- Posterior

The graphs illustrate the concepts of prior, likelihood function, and posterior distributions in Bayesian inference.
Bayesian Bernoulli

\[ p(\mu|a_0, b_0, \mathcal{D}) \propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \]
\[ = \left( \prod_{n=1}^{N} \mu^{x_n} (1 - \mu)^{1-x_n} \right) \text{Beta}(\mu|a_0, b_0) \]
\[ \propto \mu^{m+a_0-1} (1 - \mu)^{(N-m)+b_0-1} \]
\[ \propto \text{Beta}(\mu|a_N, b_N) \]

\[ a_N = a_0 + m \quad b_N = b_0 + (N - m) \]

The Beta distribution provides the conjugate prior for the Bernoulli distribution.
Properties of the Posterior

As the size of the data set, $N$, increase,

\[
\begin{align*}
a_N & \rightarrow m \\
b_N & \rightarrow N - m \\
\mathbb{E}[\mu] & = \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\text{ML}} \\
\text{var}[\mu] & = \frac{a_N b_N}{(a_N + b_N)^2(a_N + b_N + 1)} \rightarrow 0
\end{align*}
\]
Prediction under the Posterior

What is the probability that the next coin toss will land heads up?

\[ p(x = 1|a_0, b_0, \mathcal{D}) = \int_0^1 p(x = 1|\mu)p(\mu|a_0, b_0, \mathcal{D}) \, d\mu \]

\[ = \int_0^1 \mu p(\mu|a_0, b_0, \mathcal{D}) \, d\mu \]

\[ = \mathbb{E}[\mu|a_0, b_0, \mathcal{D}] \]
Bayesian Multinomial (1)

\[ p(\mu|\mathcal{D}, \alpha) \propto p(\mathcal{D}|\mu)p(\mu|\alpha) \propto \prod_{k=1}^{K} \mu_{k}^{\alpha_{k}+m_{k}-1} \]

\[ p(\mu|\mathcal{D}, \alpha) = \text{Dir}(\mu|\alpha + \mathbf{m}) \]

\[ = \frac{\Gamma(\alpha_{0} + N)}{\Gamma(\alpha_{1} + m_{1}) \cdots \Gamma(\alpha_{K} + m_{K})} \prod_{k=1}^{K} \mu_{k}^{\alpha_{k}+m_{k}-1} \]
Bayesian Multinomial (2)

\[ \alpha_k = 10^{-1} \quad \alpha_k = 10^0 \quad \alpha_k = 10^1 \]
Partitioned Gaussian Distributions (1)

\[ p(x) = \mathcal{N}(x | \mu, \Sigma) \]

\[ x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \]

\[ \Lambda \equiv \Sigma^{-1} \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \]
Let $x \sim N(\mu, \Lambda^{-1})$. We partition $x$ into two disjoint sets of random variables $x_a$ and $x_b$, which generates partitions in $\mu$ and $\Lambda$ as follows

$$
\begin{align*}
&x = \begin{bmatrix} x_a \\ x_b \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}
\end{align*}
\tag{1}
$$

We want to show that if $x$ is jointly Normal then $x_a$ and $x_b$ are also Normal and their conditional distributions conditioned on one another are also Normal.

Note that $p(x_a, x_b) = p(x_a \mid x_b) p(x_b) = p(x_b \mid x_a) p(x_a)$.

$$
p(x) = \frac{|\Lambda|^{1/2}}{(2\pi)^{d/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Lambda (x - \mu) \right] 
\tag{2}
$$
The exponential term in Equation 2 can be expanded for the partitioned data

\[
\left[ -\frac{1}{2} (x - \mu)^T \Lambda (x - \mu) \right] = -\frac{1}{2} (x_a^T \Lambda_{aa} x_a - 2 \mu_a^T \Lambda_{aa} x_a + \mu_a^T \Lambda_{aa} \mu_a \\
+ x_b^T \Lambda_{ba} x_a - 2 \mu_b^T \Lambda_{ba} x_a + \mu_b^T \Lambda_{ba} \mu_a \\
+ x_a^T \Lambda_{ab} x_b - 2 \mu_a^T \Lambda_{ab} x_b + \mu_a^T \Lambda_{ab} \mu_b \\
+ x_b^T \Lambda_{bb} x_b - 2 \mu_b^T \Lambda_{bb} x_b + \mu_b^T \Lambda_{bb} \mu_b)
\] (3)

We have to rearrange these terms to get the exponential terms for \( p(x_a | x_b) \) and \( p(x_b) \). In other words Equation 3 should be equivalent to
Partitioned Gaussian Distributions (4)

\[
\begin{align*}
\left[ -\frac{1}{2} (x - \mu)^T \Lambda (x - \mu) \right] &= \quad -\frac{1}{2} (x_a - \mu_{a|b})^T \Lambda_{a|b} (x_a - \mu_{a|b}) \\
&\quad -\frac{1}{2} (x_b - \mu_b)^T \Lambda_b (x_b - \mu_b) \\
&= \quad -\frac{1}{2} (x_a^T \Lambda_{a|b} x_a - 2 \mu_{a|b}^T \Lambda_{a|b} x_a + \mu_{a|b}^T \Lambda_{a|b} \mu_{a|b}) \\
&\quad -\frac{1}{2} (x_b^T \Lambda_b x_b - 2 \mu_b^T \Lambda_b x_b + \mu_b^T \Lambda_b \mu_b) 
\end{align*}
\]
Partitioned Gaussian Distributions (5)

Matching the terms quadratic in \( x_a \cdot (x_a^T \Lambda_{aa} x_a \) and \( x_a^T \Lambda_{a|b} x_a \) in Equations 3 and 4 we find that \( \Lambda_{a|b} = \Lambda_{aa} \).

Matching the terms linear in \( x_a \cdot (-2\mu_a^T \Lambda_{aa} x_a + x_b^T \Lambda_{ba} x_a - 2\mu_b^T \Lambda_{ba} x_a + x_a^T \Lambda_{ab} x_b \) and \(-2\mu_{a|b}^T \Lambda_{a|b} x_a \) we find that \(-2\mu_a^T \Lambda_{aa} + 2x_b^T \Lambda_{ba} - 2\mu_b^T \Lambda_{ba} = -2\mu_{a|b}^T \Lambda_{a|b} \) (we used the fact that \( x_a^T \Lambda_{ab} x_b = x_b^T \Lambda_{ba} x_a \)).
Partitioned Gaussian Distributions (6)

\[
\begin{align*}
\mu_{a|b}^T &= (\mu_a^T \Lambda_{aa} - x_b^T \Lambda_{ba} + \mu_b^T \Lambda_{ba}) \Lambda_{a|b}^{-1} \\
\Lambda_{a|b} &= \Lambda_{aa}
\end{align*}
\]  

(5)

The equation for \( \mu_{a|b} \) can be rewritten by replacing \( \Lambda_{a|b} \) by \( \Lambda_{aa} \) in the first equation in 5 to get

\[
\begin{align*}
\mu_{a|b}^T &= (\mu_a^T + (-x_b^T + \mu_b^T) \Lambda_{ba} \Lambda_{aa}^{-1}) \\
\end{align*}
\]  

(6)

We have matched the linear and quadratic terms in 3 and 4 but we have not matched the constant terms yet. Using the formulas for \( \mu_{a|b} \) and \( \Lambda_{a|b} \) we need to first expand out the constant term with respect to \( x_a \), i.e., \( \mu_{a|b}^T \Lambda_{a|b} \mu_{a|b} \).
\[
\mu_{a|b}^T \Lambda_{a|b} \mu_{a|b} = (\mu_a^T + (\mu_b - x_b)^T \Lambda_{ba} \Lambda_{aa}^{-1}) \Lambda_{aa} (\mu_a^T + (\mu_b - x_b)^T \Lambda_{ba} \Lambda_{aa}^{-1})^T \\
= (\mu_a^T \Lambda_{aa} + (\mu_b - x_b)^T \Lambda_{ba}) (\mu_a^T + (\mu_b - x_b)^T \Lambda_{ba} \Lambda_{aa}^{-1})^T \\
= \mu_a^T \Lambda_{aa} \mu_a + (\mu_b - x_b)^T \Lambda_{ba} \mu_a \\
+ \mu_a^T \Lambda_{aa} \Lambda_{aa}^{-1} \Lambda_{ba} (\mu_b - x_b) \\
+ (\mu_b - x_b)^T \Lambda_{ba} \Lambda_{aa}^{-1} \Lambda_{ba} (\mu_b - x_b) \\
= \mu_a^T \Lambda_{aa} \mu_a + \mu_b^T \Lambda_{ba} \mu_a - x_b^T \Lambda_{ba} \mu_a \\
+ \mu_a^T \Lambda_{ba} \mu_b - \mu_a^T \Lambda_{ba} x_b \\
+ (\mu_b - x_b)^T \Lambda_{ba} \Lambda_{aa}^{-1} \Lambda_{ba} (\mu_b - x_b)
\]
Partitioned Gaussian Distributions (8)

Note that the first five terms in the last equation in (7) also exists in (3) but the last term $(\mu_b - x_b)^T \Lambda_{ba} \Lambda_{aa}^{-1} \Lambda_{ba} (\mu_b - x_b)$ does not exist. If we add and subtract this term in 3 we can rewrite that equation as follows:

$$\left[ -\frac{1}{2} (x - \mu)^T \Lambda (x - \mu) \right] = (x_a - \mu_{a|b})^T \Lambda_{a|b} (x_a - \mu_{a|b})$$

$$+ (x_b - \mu_b)^T \Lambda_{bb} (x_b - \mu_b)$$

$$- (x_b - \mu_b)^T \Lambda_{ba} \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b)$$

(8)

or

$$\left[ -\frac{1}{2} (x - \mu)^T \Lambda (x - \mu) \right] = (x_a - \mu_{a|b})^T \Lambda_{a|b} (x_a - \mu_{a|b})$$

$$+ (x_b - \mu_b)^T (\Lambda_{bb} - \Lambda_{ba} \Lambda_{aa}^{-1} \Lambda_{ab}) (x_b - \mu_b)$$

(9)
If we match the second term in (9) with the second term in (4) we find:

\[
\begin{align*}
\mu_b & = \mu_b \\
\Lambda_b & = (\Lambda_{bb} - \Lambda_{ba} \Lambda_{aa}^{-1} \Lambda_{ab})
\end{align*}
\]  

Going back to equation (2) and computing the determinant of \( \Lambda \) in the block diagonal format we see that \( \det(\Lambda) = \det(\Lambda_{aa}) \det(\Lambda_{bb} - \Lambda_{ba} \Lambda_{aa}^{-1} \Lambda_{ab}) \), which indicates that equation (2) was indeed a product of two Normal distributions, i.e. \( p(x_a, x_b) = p(x_a \mid x_b) p(x_b) \) with the mean vectors and covariance matrices defined as in (5) and (10).
Partitioned Conditionals and Marginals

\[ x_b = 0.7 \]

\[ p(x_a, x_b) \]

\[ p(x_a | x_b = 0.7) \]

\[ p(x_a) \]
Bayes’ Theorem for Gaussian Variables

Given

\[ p(x) = \mathcal{N}(x|\mu, \Lambda^{-1}) \]
\[ p(y|x) = \mathcal{N}(y|Ax + b, L^{-1}) \]

we have

\[ p(y) = \mathcal{N}(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^T) \]
\[ p(x|y) = \mathcal{N}(x|\Sigma\{A^T L(y - b) + \Lambda\mu\}, \Sigma) \]

where

\[ \Sigma = (\Lambda + A^T L A)^{-1} \]