DATA11002

# Introduction to Machine Learning

Lecturer: Teemu Roos
TAs: Ville Hyvönen and Janne Leppä-aho

Department of Computer Science
University of Helsinki

(based in part on material by Patrik Hoyer and Jyrki Kivinen)

November 2nd–December 15th 2017

# Classification: Gaussian classifiers and NB

# Classification with Bayes

- Given an instance $\mathbf{x} = (x_1, \ldots, x_p)$, and any class value $c \in \{1, \ldots, k\}$, Bayes theorem gives us

$$P(Y = c \mid X = \mathbf{x}) = \frac{P(X = \mathbf{x} \mid Y = c)P(Y = c)}{\sum_{c'=1}^{k} P(X = \mathbf{x} \mid Y = c')P(Y = c')}$$

- The basic version of naive Bayes then predicts class $c$ with maximum posterior probability (MAP):

$$\hat{c}(\mathbf{x}) = \arg\max_{c} P(Y = c \mid X = \mathbf{x})$$

- Probabilistic predictions are obtained directly from $P(Y = c \mid X = \mathbf{x})$

# Classification with Bayes (2)

- Since the denominator $\sum_{c'=1}^{k} P(X = \mathbf{x} \mid Y = c')P(Y = c')$ does not depend on $c$, the MAP classification is the same as

$$\hat{c}(\mathbf{x}) = \arg\max_{c} P(X = \mathbf{x} \mid Y = c)P(Y = c)$$

- If the class prior $P(Y)$ is uniform, this simplifies to maximum likelihood (ML) prediction

$$\hat{c}(\mathbf{x}) = \arg\max_{c} P(X = \mathbf{x} \mid Y = c)$$

- The question becomes: where do we get $P(X = \mathbf{x} \mid Y = c)$ from?

# Gaussian Naive Bayes

- Assume that we have $p$ input features, $X_1, \ldots, X_p$.
- The **naive Bayes** assumption is that input features are conditionally independent given class:

$$P(X_1, \ldots, X_p \mid Y) = P(X_1 \mid Y) \ldots P(X_p \mid Y)$$

- Thus for a feature vector $\mathbf{x}$, we have

$$\begin{aligned} p(\mathbf{x} \mid Y = +1) &= p(X_1 = x_1 \mid Y = +1) \ldots p(X_p = x_p \mid Y = +1) \\ p(\mathbf{x} \mid Y = -1) &= p(X_1 = x_1 \mid Y = -1) \ldots p(X_p = x_p \mid Y = -1) \end{aligned}$$

- In the Gaussian naive Bayes model, we let $p(X_j = x \mid Y = c)$ be independent univariate Gaussians for each feature $X_j$ and class $c$

# Conditional independence

- Classical example used to illustrate conditional independence (and also difference between correlation and causation) is correlation between ice cream sales and drowning deaths
- During sunny and warm weather people tend to both eat ice cream and go boating, swimming etc. which increases chances of drowning
- Hence, there is positive correlation between ice cream sales and number of drownings on a given day
- However, if we already know what the weather actually was, then knowing how much ice cream was sold does not help us predict drowning
- Hence, ice cream sales and drownings are *conditionally* independent given weather

# Classification with Bayes (3)

- Substituting the product formula for $P(X = \mathbf{x} \mid Y = c)$ in the Bayes formula yields:

$$\hat{c}(\mathbf{x}) = \arg\max_c \left[ P(Y = c) \cdot \prod_{j=1}^{p} P(X_j = x_j \mid Y = c) \right]$$

- Alternative representation: Taking logs doesn't change the maximum, so

$$\hat{c}(\mathbf{x}) = \arg\max_c \left[ \log P(Y = c) + \sum_{j=1}^{p} \log P(X_j = x_j \mid Y = c) \right]$$

- The Gaussian assumption on $p(X_j = x_j \mid Y = c)$ leads to (a special case of) LDA/QDA! (Exercise 2.1)

# Estimating parameters

- Assume now that we have training data with positive examples $Tr^+$ and negative examples $Tr^-$
- Using maximum likelihood estimates for parameters, we get $p(X_j = x \mid Y = +1) = \mathcal{N}(x \mid \hat{\mu}_{+,j}, \hat{\sigma}^2_{+,j})$ where

$$\begin{aligned} \hat{\mu}_{+,j} &= \frac{1}{Tr^+} \sum_{x \in Tr^+} x_j \\ \hat{\sigma}^2_{+,j} &= \frac{1}{Tr^+} \sum_{x \in Tr^+} (x_j - \hat{\mu}_{+,j})^2 \end{aligned}$$

and similarly for negative examples

# Discrete Naive Bayes

- Assume now that we have $p$ **categorical** input features $X_1, \ldots, X_p$ where the possible values for $X_j$ are $\{1, \ldots, q_j\}$ for some (small) number $q_j$ of distinct values
- There are $|\mathcal{X}| = \prod_{j=1}^{p} q_j$ possible inputs we may need to classify
- Without the naive Bayes assumption, in order to determine an arbitrary distribution over $\mathcal{X}$, or an arbitrary conditional distribution $P(Y \mid X)$, we would need $|\mathcal{X}| - 1$ parameters (since probabilities sum to one but can otherwise be chosen freely to each $x \in \mathcal{X}$)
- In many realistic scenarios, $|\mathcal{X}|$ is much more than the sample size, so learning such a distribution is out of the question

# Naive Bayes classifier (2)

- Let's again make the naive Bayes assumption that input features are conditionally independent given class:

$$P(X_1, \ldots, X_p \mid Y) = P(X_1 \mid Y) \ldots P(X_p \mid Y)$$

- Each $P(X_i \mid Y)$ is determined by $q_i - 1$ (free) parameters
- For $k$ classes, the number of parameters is $k \sum_{j=1}^{p} (q_i - 1) \ll k(\prod_{j=1}^{p} q_i - 1)$

# Learning a naive Bayes model

- Assume there are $k$ classes $1, \ldots, k$ and $p$ input features where for $j = 1, \ldots, p$ feature $X_j$ has range $\{ 1, \ldots, q_j \}$
- We model $P(X \mid Y = c)$ separately for each class $c$ and feature $X \in \{ X_1, \ldots, X_p \}$:
  - For each $c \leq k$, $j \leq d$, and $x \leq q_j$, let $n_{c,j,x}$ be the number of examples in the training data in class $c$ with feature value $X_j = x$, and $n_c = \sum_{x=1}^{q_j} n_{c,j,x}$
  - We estimate

$$P(X_j = x \mid Y = c) = \frac{n_{c,j,x} + m_{c,j,x}}{n_c + m_{c,j}}$$

    where $m_{c,j,x}$ is a prior pseudocount and $m_{c,j} = \sum_{x=1}^{q_j} m_{c,j,x}$
  - Usual choices for pseudocounts are $m_{c,j,x} = 0$ (maximum likelihood), $m_{c,j,x} = 1$ (Laplace smoothing), $m_{c,j,x} = 1/2$ (Krichevsky-Trofimov), and $m_{c,j,x} = 1/3$ (the "Jääsaari method")

# About naive Bayes assumption

- The assumption that features are independent conditioned on class is
  - very strong
  - often quite untrue
- Therefore in particular the probabilities produced by a naive Bayes model should not be trusted too much
- However the classification performance (zero–one loss) of naive Bayes is often quite hard to beat in practice
- An informal justification for using naive Bayes is that often the data are collected in a way that aims to ensure (approximate) conditional independence
  - for example, in medical diagnosis, obtaining each feature requires that we carry out a test: it makes no sense to measure temperature from both armpits, or other redundant variables that we know to be strongly dependent (given the class)

# Probabilistic models: summary

- Generative probabilistic models involve modelling both $P(X \mid Y = c)$ and $P(Y = c)$ for different classes $c$
- Important tools for this include
  - multivariate Gaussians (LDA, QDA): very important overall in statistics and machine learning, important to be familiar with them
  - Naive Bayes: especially discrete NB commonly used in practice, important to understand its uses and limitations
- Discriminative probabilistic learning aims directly at $P(Y = c \mid X)$.
  - Logistic regression is a good example

# Probabilistic models in the textbook

- We have more or less covered Sec. 4 ("Classification") except pages 145–149 (class-specific accuracy, ROC curves), including **logistic regression, LDA,** and **QDA**
- In addition, we discussed **Naive Bayes** which is required for this course (and the exam) but is not covered in the book at all
- Next up: **k-NN**, **decision trees**, and **SVM**