Machine Learning - Lecture 1

Introduction

15.04.2009

Bastian Leibe
RWTH Aachen
http://www.umic.rwth-aachen.de/multimedia
leibe@umic.rwth-aachen.de

Language

- Official course language will be English
  - If at least one English-speaking student is present.
  - If not... you can choose.

- However...
  - Please tell me when I’m talking too fast or when I should repeat something in German for better understanding!
  - You may at any time ask questions in German!
  - You may turn in your exercises in German.

Organization

- Lecturer
  - Prof. Bastian Leibe (leibe@umic.rwth-aachen.de)

- Assistant
  - Dennis Mitzel (mitzel@umic.rwth-aachen.de)

- Course webpage
  - http://www.umic.rwth-aachen.de/multimedia
    → Teaching → Machine Learning
  - Slides will be made available on the webpage
  - There is also an L2P electronic repository

  Please subscribe to the lecture on the Campus system!
  - Important to get email announcements and L2P access!

Exercises and Supplementary Material

- Exercises
  - Typically 1 exercise sheet every 2 weeks
  - Pen & paper and Matlab based exercises
  - Hands-on experience with the algorithms from the lecture.
  - Send your solutions to Dennis the night before the exercise class.

- Supplementary material
  - Research papers and book chapters
  - Will be provided on the webpage.
  - Additional material will be given out if you need to take a V4 exam (Vertiefungslinie Diplom).

Textbooks

- Most lecture topics will be covered in Bishop’s book.
- Some additional topics can be found in Duda & Hart.

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006

R.O. Duda, P.E. Hart, D.G. Stork
Pattern Classification
2nd Ed., Wiley-Interscience, 2000

- Research papers will be given out for some topics.
  - Tutorials and deeper introductions.
  - Application papers
How to Find Us

- Office:
  - UMIC Research Centre
  - Mies-van-der-Rohe-Strasse 15, room 124

- Open door policy (until further notice)
  - If you have questions to the lecture, come to Dennis or me.
  - Send us an email before to confirm a time slot.

  Questions are welcome!

Machine Learning

- Goal
  - Machines that learn to perform a task from experience

- Why?
  - Crucial component of every intelligent/autonomous system
  - Important for a system’s adaptability
  - Important for a system’s generalization capabilities
  - Attempt to understand human learning

Machine Learning: Core Questions

- Learning to perform a task from experience

- Task
  - Can often be expressed through a mathematical function
  \[ y = f(x; w) \]
  - \( x \): Input
  - \( y \): Output
  - \( w \): Parameter (this is what is “learned”)

- Classification vs. Regression
  - Regression: continuous \( y \)
  - Classification: discrete \( y \)
    - E.g. class membership, sometimes also posterior probability

Example: Regression

- Automatic control of a vehicle
Examples: Classification

- Email filtering \( x \in [a-z]^* \rightarrow y \in \{\text{important, spam}\} \)
- Character recognition
- Speech recognition

Machine Learning: Core Problems

- Input \( x \):
  \( \rightarrow y \in \{\text{ahh, eeh, ... ahh}\} \)
- Features
  - Invariance to irrelevant input variations
  - Selection of the “right” features is crucial
  - Encoding and use of “domain knowledge”
- Curse of dimensionality
  - Complexity increases exponentially with number of dimensions.

Machine Learning: Core Questions

- Learning to perform a task from experience
  - Performance: “99% correct classification”
    - Of what???
    - Characters? Words? Sentences?
    - Speaker/writer independent?
    - Over what data set?
    - ...
  - “The car drives without human intervention 99% of the time on country roads”

Machine Learning: Core Questions

- Learning to perform a task from experience
  - Performance measure:
    - Typically one number
    - % correctly classified letters
    - Average driving distance (until crash...)
    - % games won
    - % correctly recognized words, sentences, answers
  - Generalization performance
    - Training vs. test
    - “All” data

Machine Learning: Core Questions

- Learning to perform a task from experience
  - Performance measure: more subtle problem
    - Also necessary to compare partially correct outputs.
    - How do we weight different kinds of errors?
    - Example: L2 norm
  - What data is available?
    - Data with labels: supervised learning
      - Images / speech with target labels
      - Car sensor data with target steering signal
    - Data without labels: unsupervised learning
      - Automatic clustering of sounds and phonemes
      - Automatic clustering of web sites
    - Some data with, some without labels: semi-supervised learning
    - No examples: learning by doing
    - Feedback/rewards: reinforcement learning
**Machine Learning: Core Questions**

- $y = f(x; w)$
  - $w$ characterizes the family of functions
  - $w$ indexes the space of hypotheses
  - $w$: vector, connection matrix, graph, ...

**Learning to perform a task from experience**

- Learning
  - Most often learning = optimization
  - Search in hypothesis space
  - Search for the “best” function / model parameter $w$
    - I.e. maximize $y = f(x; w)$ w.r.t. the performance measure

**Course Outline**

- Fundamentals (2 weeks)
  - Bayes Decision Theory
  - Probability Density Estimation
- Discriminative Approaches (5 weeks)
  - Linear Discriminant Functions
  - Statistical Learning Theory & SVMs
  - Boosting, Decision Trees
- Generative Models (5 weeks)
  - Bayesian Networks
  - Markov Random Fields
- Regression Problems (2 weeks)
  - Gaussian Processes

**Topics of This Lecture**

- (Re-)view: Probability Theory
  - Probabilities
  - Probability densities
  - Expectations and covariances
- Bayes Decision Theory
  - Basic concepts
  - Minimizing the misclassification rate
  - Minimizing the expected loss
  - Discriminant functions

**Probability Theory**

"Probability theory is nothing but common sense reduced to calculation."  
Pierre-Simon de Laplace, 1749-1827

- Example: apples and oranges
  - We have two boxes to pick from.
  - Each box contains both types of fruit.
  - What is the probability of picking an apple?
- Formalization
  - Let $B \in \{r, b\}$ be a random variable for the box we pick.
  - Let $F \in \{a, o\}$ be a random variable for the type of fruit we get.
  - Suppose we pick the red box 40% of the time. We write this as $p(B = r) = 0.4$ and $p(B = b) = 0.6$.
  - The probability of picking an apple given a choice for the box is $p(F = a | B = r) = 0.25$ and $p(F = a | B = b) = 0.75$.
  - What is the probability of picking an apple? $p(F = a)$ = ?
Probability Theory

- More general case
  - Consider two random variables $X \in \{x_i\}$ and $Y \in \{y_j\}$
  - Consider $N$ trials and let $n_{ij} = \#\{X = x_i \land Y = y_j\}$
  - Then we can derive
    - Joint probability $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$
    - Marginal probability $p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$
    - Conditional probability $p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$

- Rules of probability
  - Sum rule $p(X = x_i) = \frac{1}{N} \sum_j \frac{1}{N} \sum_j n_{ij} = \sum_j p(X = x_i, Y = y_j)$
  - Product rule $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = p(Y = y_j | X = x_i) p(X = x_i)$

The Rules of Probability

- Thus we have
  - Sum Rule $p(X) = \sum_Y p(X, Y)$
  - Product Rule $p(X, Y) = p(Y | X)p(X)$

- From those, we can derive
  - Bayes’ Theorem $p(Y | X) = \frac{p(X | Y)p(Y)}{p(X)}$ where $p(X) = \sum_Y p(X, Y) p(Y)$

Probability Densities

- Probabilities over continuous variables are defined over their probability density function (pdf) $p(x)$.
- The probability that $x$ lies in the interval $(a, b)$ is given by the cumulative distribution function $P(z) = \int_a^z p(x) \, dx$

Expectations

- The average value of some function $f(x)$ under a probability distribution $p(x)$ is called its expectation
  - Discrete case $\mathbb{E}[f] = \sum_x p(x) f(x)$
  - Continuous case $\mathbb{E}[f] = \int f(x) p(x) \, dx$

- If we have a finite number $N$ of samples drawn from a pdf, then the expectation can be approximated by $\mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)$

- We can also consider a conditional expectation $\mathbb{E}_X[f | y] = \sum_x p(x | y) f(x)$

Variances and Covariances

- The variance provides a measure how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$.
  - Variance $\text{Var}[f] = \mathbb{E}[\{f(x) - \mathbb{E}[f(x)]\}^2] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$

- For two random variables $x$ and $y$, the covariance is defined by $\text{cov}[x, y] = \mathbb{E}_x \mathbb{E}_x [\{x - \mathbb{E}[x]\} \{y - \mathbb{E}[y]\}]$.

- If $x$ and $y$ are vectors, the result is a covariance matrix $\text{cov}[x, y] = \mathbb{E}_x [\{x - \mathbb{E}[x]\} \{y - \math{E}[y]\}]$.
Bayes Decision Theory

Example: handwritten character recognition

Goal:
Classify a new letter such that the probability of misclassification is minimized.

Concept 1: Priors (a priori probabilities)
What we can tell about the probability before seeing the data.
Example:

\[
p(C_1) = 0.75 \quad p(C_2) = 0.25
\]

In general:
\[
\sum_k p(C_k) = 1
\]

Concept 2: Conditional probabilities
Let \( x \) be a feature vector.
\( x \) measures/describes certain properties of the input.
E.g., number of black pixels, aspect ratio, ...
\( p(x|C_k) \) describes its likelihood for class \( C_k \).

Example:

\[
p(x|a) \quad p(x|b)
\]

Question:
Which class?
The decision should be 'a' here.

Question:
Which class?
Since \( p(x|a) \) is much smaller than \( p(x|b) \), the decision should be 'b' here.
Bayes Decision Theory

- **Example:**
  
  ![Graph showing probability distributions](image)

  - **Question:**
    - Which class?
    - Remember that \( p(a) = 0.75 \) and \( p(b) = 0.25 \), i.e., the decision should be again ‘a’.
    - How can we formalize this?

Bayes Decision Theory

- **Bayes Decision Theory**
  
  ![Bayes Decision Theory diagram](image)

  - **Goal:** Minimize the probability of a misclassification

  \[
  p(\text{misclass}) = p(x \in \mathcal{R}_1, C_1) + p(x \in \mathcal{R}_2, C_1)
  = \int_{\mathcal{R}_1} p(x, C_0) \, dx + \int_{\mathcal{R}_2} p(x, C_1) \, dx
  = \int_{\mathcal{R}_1} p(C_0|x) p(x) \, dx + \int_{\mathcal{R}_2} p(C_1|x) p(x) \, dx
  \]

  - **Concept 3: Posterior probabilities**

  - We are typically interested in the a posteriori probability, i.e., the probability of class \( C_i \) given the measurement vector \( x \).

  - **Bayes’ Theorem:**

  \[
  p(C_i | x) = \frac{p(x | C_i) p(C_i)}{\sum_j p(x | C_j) p(C_j)}
  \]

  - **Interpretation**

  \[
  \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalization Factor}}
  \]

Bayesian Decision Theory

- **Generalization to More Than 2 Classes**

  - Decide for class \( i \) whenever it has the greatest posterior probability of all classes:

  \[
  p(C_i | x) > p(C_j | x) \quad \forall j \neq k
  \]

  \[
  p(x | C_k) p(C_k) > p(x | C_j) p(C_j) \quad \forall j \neq k
  \]

  - **Likelihood-ratio test**

  \[
  \frac{p(x | C_k)}{p(x | C_j)} > \frac{p(C_j)}{p(C_k)} \quad \forall j \neq k
  \]
Classifying with Loss Functions

- Generalization to decisions with a **loss function**
  - Differentiate between the possible decisions and the possible true classes.
  - Example: medical diagnosis
    - Decisions: sick or healthy (or: further examination necessary)
  - Classes: patient is sick or healthy
- The cost may be asymmetric:
  - \( \text{loss}(\text{decision} = \text{healthy} | \text{patient} = \text{sick}) > > \text{loss}(\text{decision} = \text{sick} | \text{patient} = \text{healthy}) \)

Minimizing the Expected Loss

- In general, we can formalize this by introducing a loss matrix \( L_{kj} \)
  \[ L_{kj} = \text{loss for decision } C_j \text{ if truth is } C_k. \]
- Example: cancer diagnosis
  \[ L_{\text{cancer diagnosis}} = \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \]

Minimizing the Expected Loss

- Optimal solution is the one that minimizes the loss.
  - But: loss function depends on the true class, which is unknown.
- Solution: **Minimize the expected loss**
  \[ \mathbb{E}[L] = \sum_k \sum_j \int \mathbb{E}[L_{kj}] \, dx \]
  - This can be done by choosing the regions \( R_j \) such that
  \[ \mathbb{E}[L] = \sum_k \mathbb{E}[L_{kj}] \]
  which is easy to do once we know the posterior class probabilities \( p(C_k | x) \).

Minimizing the Expected Loss

- Example:
  - 2 Classes: \( C_1, C_2 \)
  - 2 Decision: \( \alpha_1, \alpha_2 \)
  - Loss function: \( L(\alpha_j | C_i) = L_{kj} \)
  - Expected loss (or risk) \( \mathbb{E}[L(\alpha_j | C_i)] = R(\alpha_j | x) \)
  for the two decisions:
  \[ R(\alpha_2 | x) = R(\alpha_1 | x) = L_{11} p(C_1 | x) + L_{21} p(C_2 | x) \]
  - Goal: Decide such that expected loss is minimized
  - i.e. decide \( \alpha \) if \( R(\alpha_2 | x) > R(\alpha_1 | x) \)

Minimizing the Expected Loss

- Example:
  - 2 Classes: \( C_1, C_2 \)
  - 2 Decision: \( \alpha_1, \alpha_2 \)
  - Loss function: \( L(\alpha_j | C_i) = L_{kj} \)
  - Expected loss (or risk) \( \mathbb{E}[L(\alpha_j | C_i)] = R(\alpha_j | x) \)
  for the two decisions:
  \[ R(\alpha_2 | x) = R(\alpha_1 | x) = L_{11} p(C_1 | x) + L_{21} p(C_2 | x) \]
  - Goal: Decide such that expected loss is minimized
  - i.e. decide \( \alpha \) if \( R(\alpha_2 | x) > R(\alpha_1 | x) \)
The Reject Option

- Classification errors arise from regions where the largest posterior probability \( p(C_i | x) \) is significantly less than 1.
  - These are the regions where we are relatively uncertain about class membership.
  - For some applications, it may be better to reject the automatic decision entirely in such a case and e.g. consult a human expert.

Discriminant Functions

- Formulate classification in terms of comparisons
  - Discriminant functions
    \( y_1(x), \ldots, y_K(x) \)
  - Classify \( x \) as class \( C_k \) if
    \( y_k(x) > y_j(x) \quad \forall j \neq k \)

- Examples (Bayes Decision Theory)
  \[
  y_k(x) = p(C_k | x) \\
  y_k(x) = p(x | C_k) p(C_k) \\
  y_k(x) = \log p(x | C_k) + \log p(C_k)
  \]

Next Lectures...

- Ways how to estimate the probability densities \( p(x | C_k) \)
  - Non-parametric methods
    - Histograms
    - k-Nearest Neighbor
    - Kernel Density Estimation
  - Parametric methods
    - Gaussian distribution
    - Mixtures of Gaussians
  - Discriminant functions
    - Linear discriminants
    - Support vector machines

References and Further Reading

- More information, including a short review of Probability theory and a good introduction in Bayes Decision Theory can be found in Chapters 1.1, 1.2 and 1.5 of
  
  Christopher M. Bishop
  Pattern Recognition and Machine Learning
  Springer, 2006

Image source: C.M. Bishop, 2006