Statistical Machine Learning

Lecture 1 – Introduction and math preliminaries

UPPSALA
UNIVERSITET

Jens Sjölund
https://jsjol.github.io/
Department of Information Technology
Uppsala University

Course webpage

UPPSALA
UNIVERSITET

What is the course about?

Machine learning

"Machine learning is about learning, reasoning and acting based on data."

Three cornerstones:

1. Data.
2. Mathematical model.
3. Learning algorithm.

UPPSALA UNIVERSITET

Focus of this course

Supervised machine learning

Methods for learning a predictive **model** $f(\mathbf{x}) = y$ of the relationship between

- features of possible relevance $\mathbf{x}$ (**input**)
- outcome of interest y (**output**)

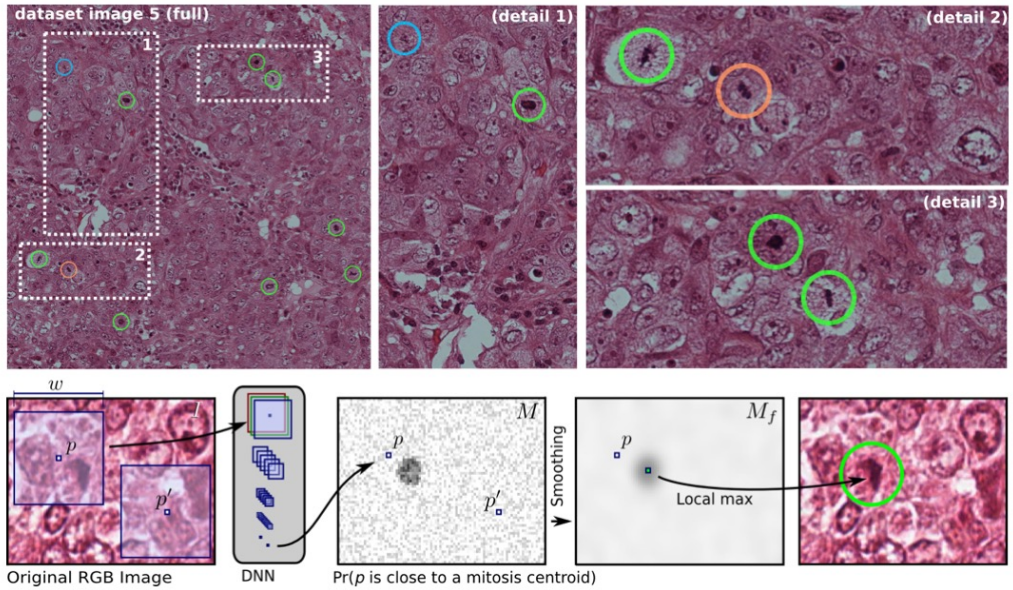
from observed **training data**

$$\mathcal{T} \stackrel{\text{def}}{=} \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}.$$

UPPSALA UNIVERSITET

ex) Cancer diagnosis

Systems for detecting cell divisions (mitosis) in histology images can be used to improve (or automate) cancer diagnosis.



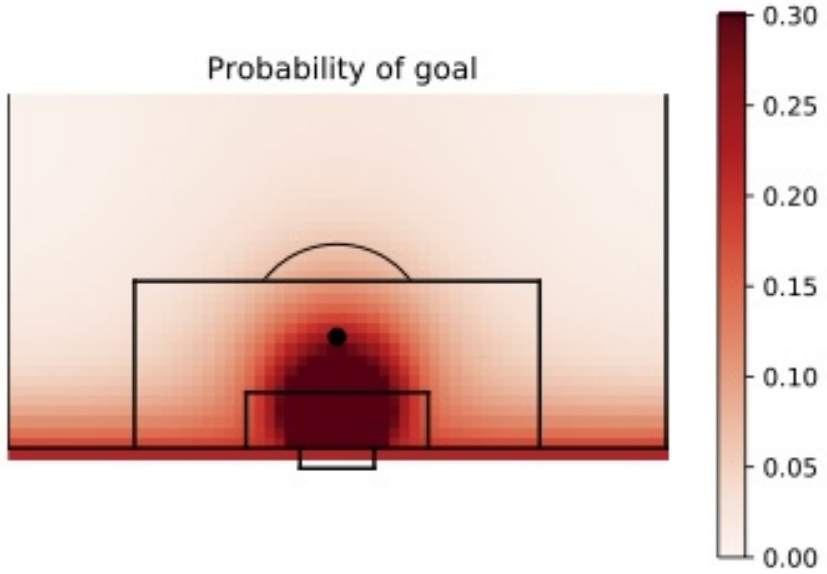
- Learn a model with
  input $\mathbf{x}$ = RGB histology image (pixel values)
  output y = number and locations (in the image) of mitosis detections
- Training data: Histology images labeled by experts.
- Uses deep neural networks to model f (Lectures 7 and 8).

D. C. Cireşan, A. Giusti, L. M. Gambardella and J. Schmidhuber. **Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks**. In *Medical Image Computing and Computer Assisted Intervention*, 411-418, 2013.

ex) Goals in football

- Predicting the probability a shot is a goal
- Learn a model with

  input $\mathbf{x} =$ location, header/foot, left/right foot, from cross/pass ... ...
  output $y =$ goal
- Logistic regression (Lecture 3) and Random forest models (Lecture 6).



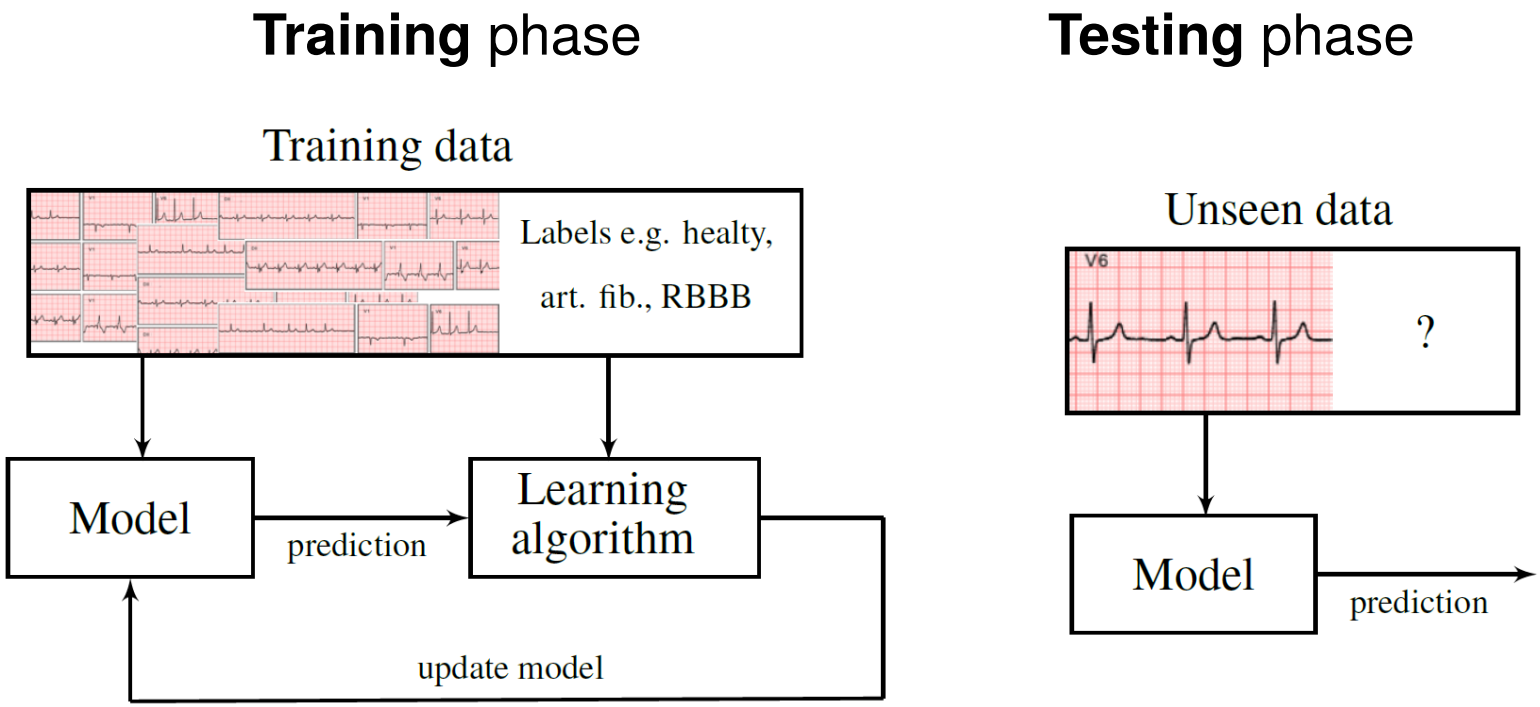
Statistical machine learning

Why the word "statistical" in the course title?

- Probability theory is used to define the models.
- Statistical tools are used to learn the models from training data.

Allows us to reason about the ***uncertainties*** in the data, models, predictions, etc.!

Training and testing



Key assumption: The unseen testing data comes from the same probability distribution as the training data.

Quantitative and qualitative variables

Both input variables ($\mathbf{x}$) and output variables (y) can be either **quantitative** or **qualitative**.

- **Quantitative** variables take on numerical values (real numbers, integer values, . . .).
- **Qualitative** variables take on values in one of K distinct classes, e.g. “true or false”, “disease type A, B or C”.

We will mostly use integer coding (like $\{1, 2, 3, 4\}$ or $\{0, 1\}$), but the coding (or labeling) of qualitative variables is arbitrary and unordered.

Regression vs. classification

We will distinguish between two types of problems:
regression and **classification**

Regression is when the output y is quantitative, e.g.

- Climate models ($y =$ "increase in global temperature")
- Socio-economic models ($y =$ "change in GDP", "happiness")
- Social media engagement ($y =$ "watch time of Youtube videos")

Classification is when the output y is qualitative, e.g.

- Spam filters ($y \in \{\texttt{spam}, \texttt{proper email}\}$).
- Face recognition, e.g. on your phone ($y \in \{\texttt{match}, \texttt{no match}\}$).

Models

Different machine learning methods use different **models** for the function f (even if the task at hand may be the same).

"All models are wrong, but some are useful" — *George Box*

Some methods use more ***flexible models*** than others, capable of capturing complex dependencies between inputs and outputs.

Why use a more restrictive model instead of a flexible one?

- Flexible models tend to be harder to **interpret**.
- Flexible models tend to be less **robust**.

UPPSALA
UNIVERSITET

Course information

Course elements

- 10 lectures
- 10 exercise sessions
- 1 project (3-4 students, written report)
- 1 computer lab (4h, no report)
- Everything distributed via the course webpage

Lecture outline

1. Introduction
- Introduction to statistical thinking
- Introduction to Python
2. Basic regression, linear regression
3. Classification, logistic regression
4. Classification, LDA, QDA, k-NN
5. Generalization performance
6. Tree-based methods, bagging, boosting
7. Deep neural networks
8. Model learning as numerical optimization
9. Limitations and new directions for supervised learning
10. Summary and outlook

You can post questions that you want us to discuss in the lectures in our Padlet.

Exercise sessions

10 exercise sessions in total:

- Solve problems before the session
- Ask questions in the discussion forum. We will check them every day.
- 5 pen-and-paper sessions.
- 5 computer-based sessions (using Python).
- Feel free to use your own laptop – Python is freely available.

Each session 3 times, 2 in parallel.

A great opportunity to discuss and ask questions!

Course webpage

- All information should be available on the course webpage.
- We are 300 people trying learn about a subject. Let's help each other as much as possible!
- If you have a question that someone else might also have, please **post it in the discussion forum**.
- If you absolutely have to contact us directly, then please contact **only one of the staff** via Studium. However, if we consider the question of potential interest to others, we will restate and answer it in the discussion forum.
- Please use the helpdesks to get help with the project.

Examination

Project:

- Solved in groups of 3 or 4 students (sign-up opens soon)
- Written report (deadline: see home page)
- Peer-review: read and review another group's report (anonymously)
- Material most relevant for the project presented in lectures 3–6, **but you can start working on the solution after lecture 3**
- We offer **helpdesks** if you have questions about the project.
- Graded U/G/VG. Grade VG is for projects of notable quality.

Lab:

- 4 h computer lab, solved in groups of 2 students, graded U/G
- There will be an announcement when you can sign up
- **The preparatory exercises need to be submitted before the first lab session!**

Written exam:

- Written pen-and-paper exam, graded as U, 3, 4, or 5.
- Old exams are available on the course home page

UPPSALA UNIVERSITET

Examination

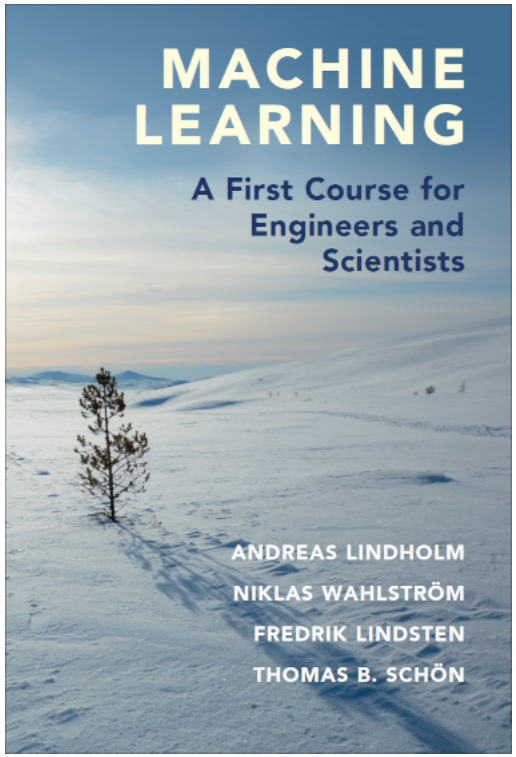
Grading:

- To pass the course you need to pass the project, lab and exam.
- Grade VG on the project will increase your course grade by one step, but not from U to 3.

		Exam grade			
		U	3	4	5
Grade on project	U	U	U	U	U
	G	U	3	4	5
	VG	U	4	5	5
		Course grade			

Secret to passing: Revise foundations and get your hands dirty with problem-solving!

Course literature



Available in print or for free via **smlbook.org**.
Possible to give feedback at book Github page.

UPPSALA
UNIVERSITET

Mathematical preliminaries

Discrete random variables

A **random variable** X can be either discrete or continuous.

- For a **discrete** random variable we write $p(x \mid \theta)$ for the probability of $X = x$ given the parameter(s) θ.
- Here, $p(x \mid \theta)$ is also called the probability mass function. It is nonnegative and sums to one, hence $0 \leq p(x \mid \theta) \leq 1$.

Example The outcome of a coin flip can be described by a random variable ($X = 1$ if heads, $X = 0$ if tails) that follows the Bernoulli distribution

$$p(x|\theta) = \begin{cases} \theta & , \quad \text{if } x = 1 \\ 1 - \theta & , \quad \text{if } x = 0 \end{cases}$$

If it's a fair coin, then $\theta = 0.5$.

Continuous random variables

A **random variable** X can be either discrete or continuous.

- For a **continuous** random variable, we write $p(x \mid \theta)$ for the probability density function (PDF).
- The PDF is nonnegative and **integrates** to unity, hence it is not bounded above in general.
- The probability is the area under the PDF, i.e. we have to integrate to get a **probability**, e.g.

$$P(X \leq x|\theta) = \int_{-\infty}^{x} p(z|\theta)dz.$$

Continuous random variables

Example If X is uniformly distributed on the interval $[1.1, 1.3]$, then

$$p(x) = \begin{cases} 5, & \text{if } x \in [1.1, 1.3] \\ 0, & \text{otherwise.} \end{cases}$$

The probability that $X \in [1.15, 1.25]$ is given by

$$P(X \in [1.15, 1.25]) = \int_{1.15}^{1.25} p(x)\, dx = 5 \cdot 0.1 = 0.5.$$

Mean and Variance

- If X is a random variable with PDF $p(x)$, then the **expected value** or **mean** of X is given by

$$\mu_X = \mathbb{E}[X] = \int x\, p(x) \mathrm{d}x.$$

- The **variance** of X is defined as

$$\mathrm{Var}[X] = \mathbb{E}[(x-\mu)^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

- The **standard deviation** of X is the square root of the variance

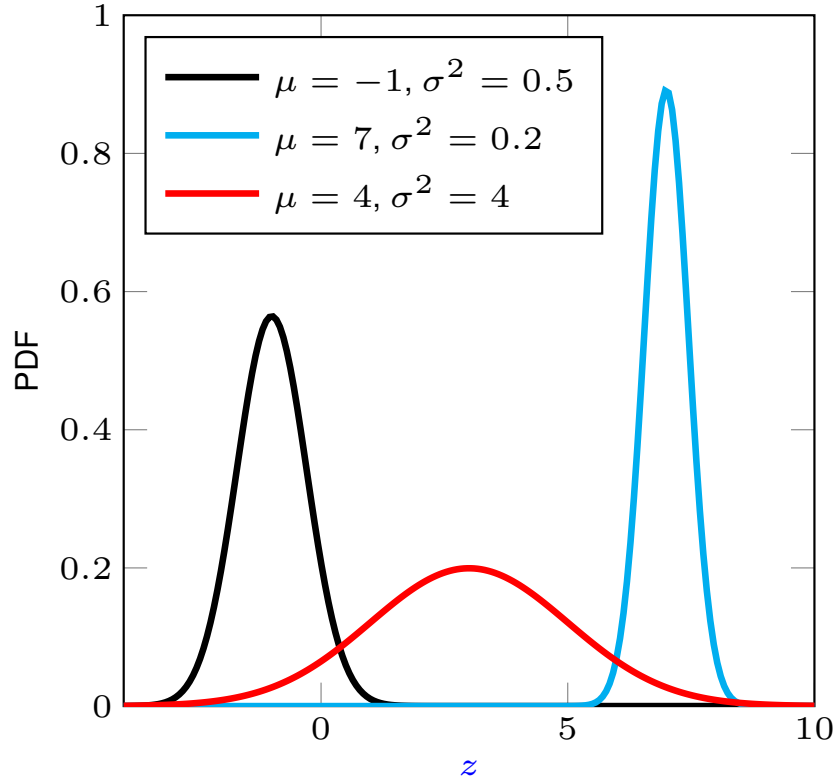
$$\sigma_X = \sqrt{V[X]}.$$

  Conversely, the variance $V[X]$ is often written as σ_X^2.

Normal (Gaussian) distribution

Probability density function for the scalar Gaussian distribution

$$p(x|\mu,\sigma^2) = \mathcal{N}(x \,|\, \mu, \sigma^2)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- μ is the mean (expected value of the distribution)
- σ is the standard deviation
- σ^2 is the variance

$X \sim \mathcal{N}(\mu, \sigma^2)$ means that X is a Gaussian random variable with mean μ and variance σ^2. $\sim$ reads “distributed according to”.

Independent random variables

Example Assume we roll two dice and denote their outcomes X_1 and X_2. This can be represented as a multivariate (bivariate) random variable $X = (X_1, X_2)$ with a corresponding **joint distribution** $p(x_1, x_2)$.
Since the die rolls are iid, $p(x_1, x_2) = p(x_1)p(x_2) = \frac{1}{36}$.

In general, if X_1 and X_2 are **independent** random variables, then the joint distribution factorizes $p(x_1, x_2) = p(x_1)p(x_2)$ and

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2].$$

In this course, we will often talk about variables that are independent and identically distributed (**iid**).

Covariance and correlation

- The **covariance** between two random variables X and Y is defined as

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^T] \\ &= \mathbb{E}[XY^T] - \mu_X \mu_Y^T.\end{aligned}$$

  If X and Y are vectors, then $\mathrm{Cov}(X, Y)$ is a matrix.

- The **correlation** between two random variables X and Y is defined as

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

- Independent variables are uncorrelated, i.e. have zero correlation, but the converse does not hold in general.

Covariance and correlation

Example Assume we roll two dice and denote their outcomes X_1 and X_2, and their sum $Y = X_1 + X_2$. What is the covariance between X_1 and Y?

$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = \sum_{k=1}^{6} k\, p(k) = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\mathbb{E}[Y] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = 7 \quad \text{(linearity of expectations)}$$

$$\mathbb{E}[X_1 Y] = \mathbb{E}[X_1(X_1 + X_2)] = E[X_1^2] + \mathbb{E}[X_1]\mathbb{E}[X_2]$$

$$= \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6} + 3.5^2 \approx 27.4$$

$$\text{Cov}(X_1, Y) = \mathbb{E}[X_1 Y] - \mathbb{E}[X_1]\mathbb{E}[Y] = 27.4 - 3.5 \cdot 7 = 2.9$$

Since $\text{Cov}(X_1, Y) \neq 0$, X_1 and Y are **not** independent.

Conditional distribution

The **conditional distribution** $p(x \mid y)$ is the probability distribution of X given that $Y = y$.

Example Assume we roll two dice and denote their outcomes X_1 and X_2, and their sum $Y = X_1 + X_2$. What is the conditional distribution of Y given that $X_1 = 3$?

Given that $X_1 = 3$, the sum $Y = 3 + X_2$. Hence,

$$p(y \mid X_1 = 3) = \begin{cases} 1/6, & \text{if } y \in \{4, \ldots, 9\} \\ 0, & \text{otherwise.} \end{cases}$$

Marginal distribution

The **marginal distribution** of a subset of random variables is the probability distribution when disregarding the other variables.

Example Assume we roll two dice and denote their outcomes X_1 and X_2, and their sum $Y = X_1 + X_2$. What is the marginal distribution of Y?

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$$p(y) = \begin{cases} 1/36, & y \in \{2, 12\} \\ 2/36, & y \in \{3, 11\} \\ 3/36, & y \in \{4, 10\} \\ 4/36, & y \in \{5, 9\} \\ 5/36, & y \in \{6, 8\} \\ 6/36, & y \in \{7\} \\ 0, & \text{otherwise.} \end{cases}$$

A few concepts to summarize lecture 1

Machine Learning: Deals with learning, reasoning and acting based on data.

Regression: Learning problem where the *output* is quantitative.

Classification: Learning problem where the *output* is qualitative.

Training and testing: We learn a predictive model on the training data and want to use on the test data.

Probability and statistics: The basis for most of machine learning.