Lecture 3

29.09.2025

Today's plan and announcements

- Review: last week's summary
- Introduction to classification
- Logistic regression

- Class reminder: we aim to have a respectful atmosphere. This means no talking/ whispering during lecture. You are welcome to ask questions if something is unclear, or speak during the break, discussion periods. But outside these times, for creating a good learning environment for everyone and keeping respect for your instructor and teaching, you should not talk/whisper/listen to audio/ watch videos. If this does not serve you, please note that attendance is not a requirement.

Review from last week, exercise

- a) You are given a dataset $\{(x^i, y^i)\}_{i=1}^3 = \{(1,2),(2,3),(3,5)\}$. You will use linear regression to predict label of a new datapoint x^{test}. Consider a linear predictor: $f(x) = b + wx$
  1. Write the linear regression loss function given the above training data.
  2. Derive the gradient of the loss function with respect to b, w
  3. Why is the loss convex in the predictor parameters b, w?
  4. Predict label of the x^{test}.

- b) You find that your prediction error was large and consider a 5-degree polynomial as predictor.
  5. Would you risk overfitting or underfitting?
  6. How would you regularize the loss?

Review from last week, solution

1a) Loss: $J(b,w) = \frac{1}{3}\sum_{i=1}^{3}\left(b + w\,x^i - y^i\right)^2 = \frac{1}{3}\left[(b + w\cdot 1 - 2)^2 + (b + w\cdot 2 - 3)^2 + (b + w\cdot 3 - 5)^2\right].$

2a) Gradient wit respect to the parameters

$$\frac{\partial J}{\partial b} = \frac{2}{3}\sum_{i=1}^{3}\left(b + w\,x^i - y^i\right), \qquad \frac{\partial J}{\partial w} = \frac{2}{3}\sum_{i=1}^{3}\left(b + w\,x^i - y^i\right)x^i.$$

3a) Convexity: $\nabla^2 J(b,w) = \frac{2}{3}\sum_{i=1}^{3}\begin{pmatrix}1 & x^i\\ x^i & (x^i)^2\end{pmatrix} = \frac{2}{3}X^\top X,$ where $X = \begin{pmatrix}1 & x^1\\ 1 & x^2\\ 1 & x^3\end{pmatrix}$, since $X^\top X$ is positive semidefinite,

the cost is convex with respect to the parameters (we already showed it abstractly for any linear regression problem)

4a) $\hat{y}^{\text{test}} = f(x^{\text{test}}) = b^* + w^*\,x^{\text{test}}$, where b^*, w^* are the unique solutions to the minimization above.

1b). High model complexity and very little data -> we risk overfitting to the training data

2b)$J^{\text{reg}}(b,w) = J(w,b) + \lambda w^2$, where λ is the regularization parameter.

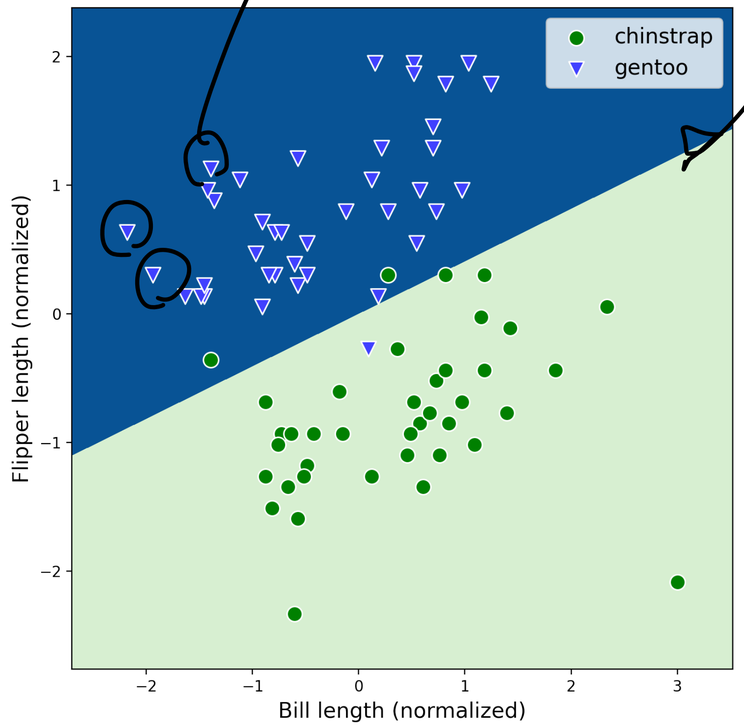
Logistic regression

classification $\{x^i, y^i\}_{i=1}^N$, $y^i \in \{1, 2, \cdots, K\}$

Classification

Palmer Penguins

	species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
0	Chinstrap	49.0	19.5	210.0	3950.0
1	Chinstrap	50.9	19.1	196.0	3550.0
2	Gentoo	42.7	13.7	208.0	3950.0
3	Chinstrap	43.5	18.1	202.0	3400.0
4	Chinstrap	49.8	17.3	198.0	3675.0



Based on dataset available here: https://www.kaggle.com/code/parulpandey/penguin-dataset-the-new-iris

Binary logistic regression

$y^i \in \{0, 1\}$

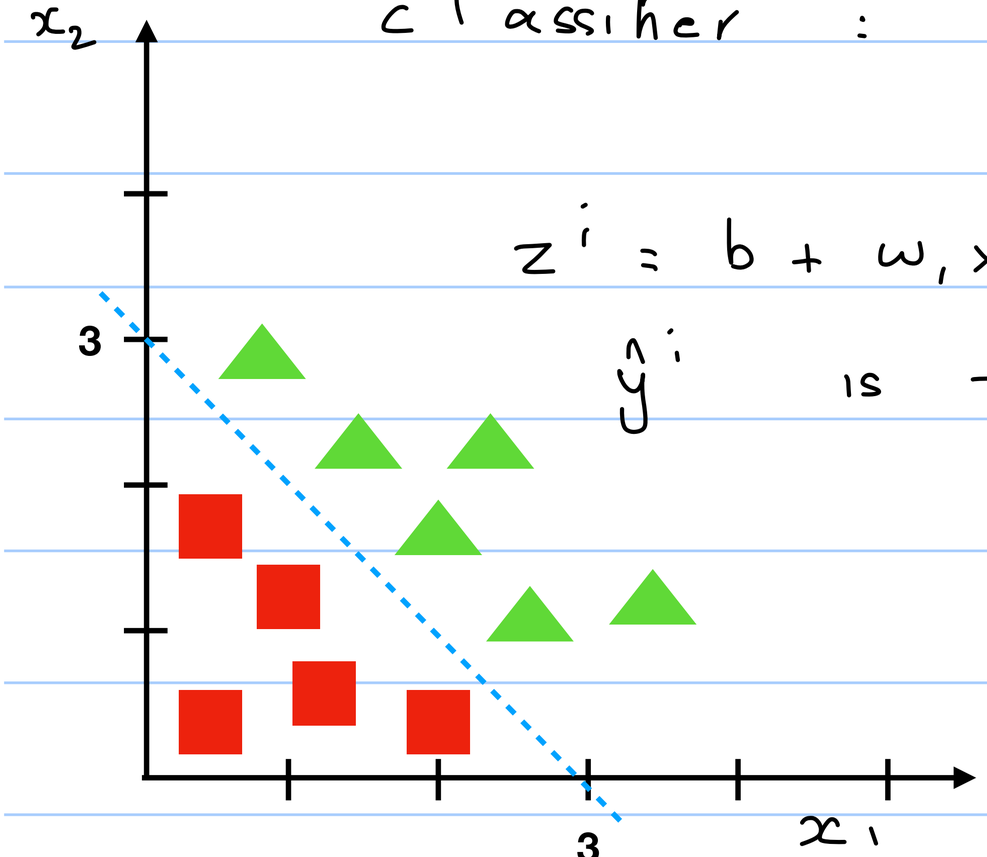
Goal: find a line/hyperplane separating the classes

$$z = b + w_1 x_1 + \cdots + w_d x_d$$

classifier : $\begin{cases} z^i \geq 0 & \hat{y}^i \text{ class } 1 \\ z^i < 0 & \hat{y}^i \text{ class } 0 \end{cases}$

$$z^i = b + w_1 x_1^i + \cdots + w_d x_d^i$$

$\hat{y}^i$ is the prediction.



Example:

$$(b, w_1, w_2) = (-3, 1, 1) \rightarrow z = x_1 + x_2 - 3$$

Predict triangle if $z \geq 0$ for a given x.

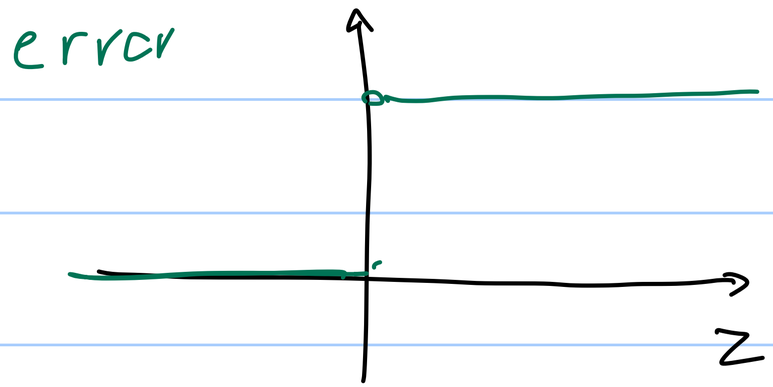
Defining loss function for binary classification

- Classifier $z = b + w_1 x_1 + \cdots + w_d x_d$, $\begin{cases} z \geq 0 \\ z < 0 \end{cases}$ class 1, $\hat{y}$ / class 0, $\hat{y}$

Dataset $\{x^i, y^i\}_{i=1}^N$

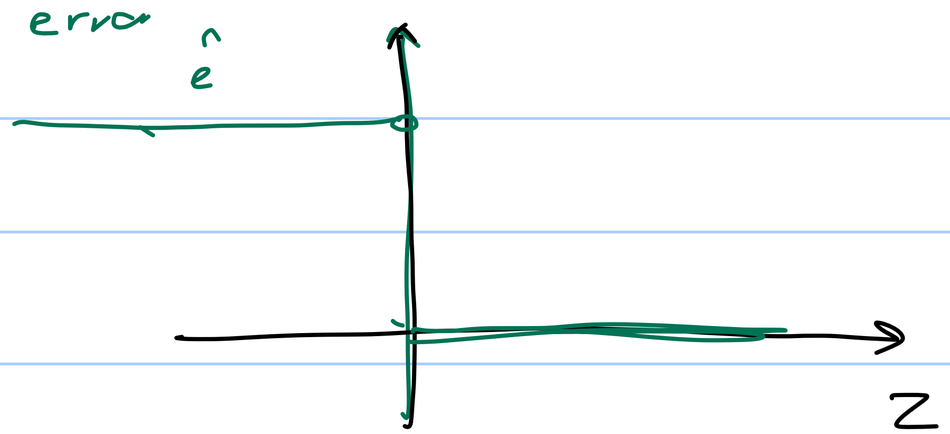
- Error

true label $y = 0$



$|\hat{y} - y|$

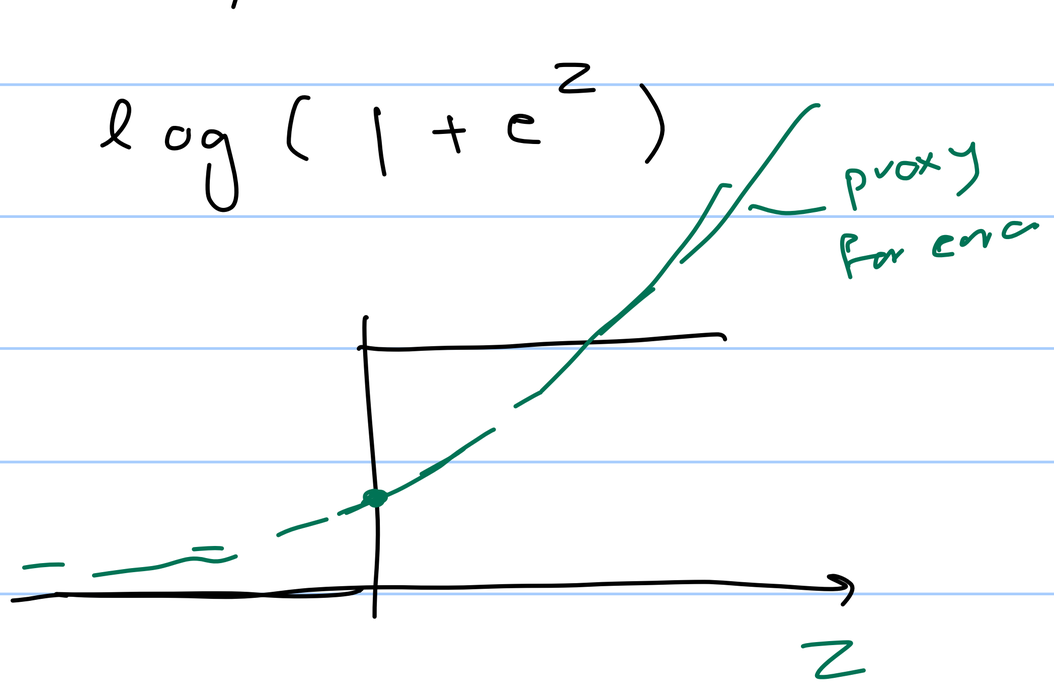
true label $y = 1$



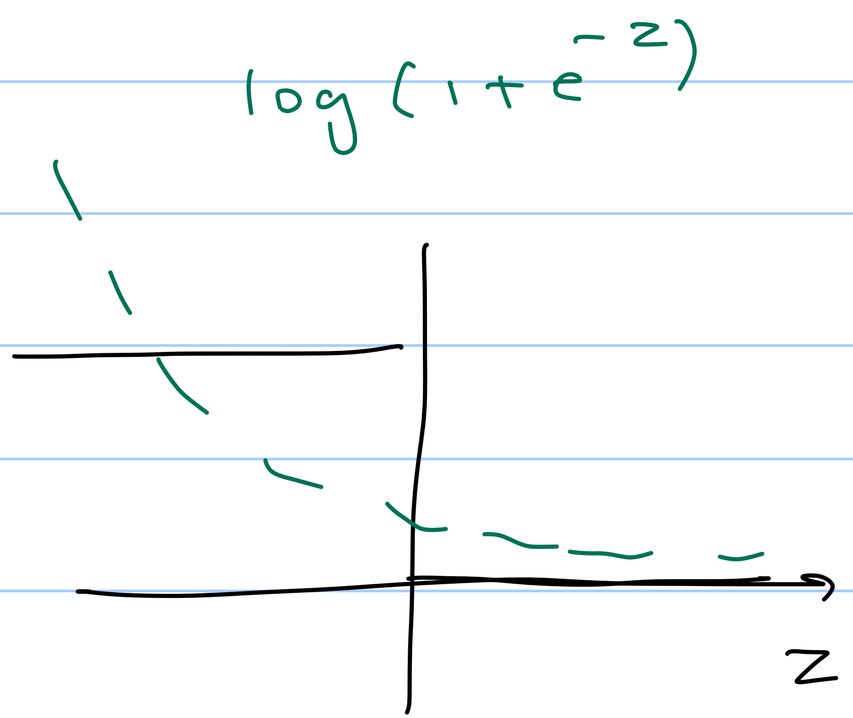
- Problem: this function is not differentiable (not even continuous). How do we optimise it?

Logistic loss function
Differentiable and convex proxy for error function

- Loss for true class 0, prediction 1



- Loss for true class 1, prediction 0



$z = b + w_1 x_1 + \cdots + w_d x_d$, need to find $(b, w_1, \cdots, w_d)$

- Why is the function differentiable? How would you verify its convexity?

Read more about logistic function/verify your gradient computation here: https://en.wikipedia.org/wiki/Logistic_function

Training - minimizing the logistic loss function

- Training data $\{x^i, y^i\}_{i=1}^N$, $x^i \in \mathbb{R}^d$, $y^i \in \{0,1\}$ (true label)
- Classifier $z^i = b + w_1 x_1^i + \cdots + w_d x_d^i$, $\hat{y}^i = \begin{cases} 1 & z^i \geq 0 \\ 0 & z^i < 0 \end{cases}$

- Logistic loss

$$J(b, w_1, \ldots, w_d) =$$

$$\frac{1}{N} \sum_{i=1}^{N} \left[\underbrace{(1 - y^i)}_{\text{is 1 iff } y^i = 0} \log\left(1 + e^{z^i}\right) + y^i \log\left(1 + e^{z^i}\right) \right] =$$

$$\frac{1}{N} \sum_{i=1}^{N} \left[1_{\{y^i = 0\}} \log\left(1 + e^{z^i}\right) + 1_{\{y^i = 1\}} \log\left(1 + e^{-z^i}\right) \right]$$

Exercise (do for exercise hour)

Consider the sigmoid function $\sigma : \mathbb{R} \to (0,1), \sigma(z) = \frac{1}{1+e^{-z}}$.

- **Compute** $\frac{d\sigma(z)}{dz}$.
- **Use chain rule to compute** $\frac{d\sigma(z(w))}{dw}$**, where** $z = w_0 + w_1 x_1 + \ldots + w_d x_d$ **and** $w = (w_0, w_1, \ldots, w_d)$
- **How would you verify that the logistic loss function is convex, based on the above?**
- **Compute the gradient of the logistic loss function with respect to the parameters** $w = (w_0, w_1, \ldots, w_d)$
- **Write the pseudo-code for finding the optimal parameters using gradient descent.**

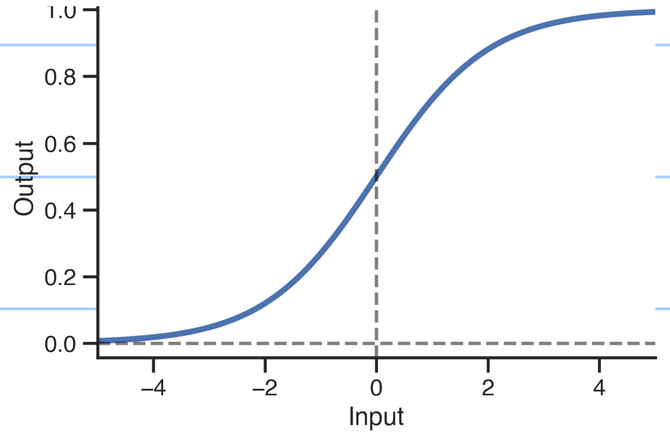
Probabilistic interpretation of logistic function

- Consider the logistic loss $$\frac{1}{N}\sum_{i=1}^{N} y^i \log(1+e^{-z^i}) + (1-y^i)\log(1+e^{z^i})$$

- Verify (simple algebra) that it's equivalent to $$\frac{-1}{N}\sum_{i=1}^{N} y^i \log(\frac{1}{1+e^{-z^i}}) + (1-y^i)\log(\frac{e^{-z^i}}{1+e^{-z^i}})$$

$\sigma(z) = \frac{1}{1+e^{-z}}$, $\sigma : \mathbb{R} \rightarrow (0,1)$

- Sigmoid function and probability of predicting class 1



Given probabilistic prediction, how can we interpret the above loss probabilistically?

To answer the question, we need to briefly review probability

Probability distribution review

- Probability distribution

is a probability distribution on a finite set

$$S = \{s_1, s_2, \ldots, s_n\}, \quad p = (p(s_1), \ldots, p(s_n)), \; p(s_j) \geq 0, \quad \sum_{j=1}^{n} p(s_j) = 1.$$

probability density function

$$S \subseteq \mathbb{R}^d, \; \forall A \subset \mathbb{R}^d, \; \int_A p(s)ds \geq 0, \; \int_S p(s)ds = 1.$$

- Example: coin flip $S = \{s_1 = \text{heads}, s_2 = \text{tails}\}$

$p = (1/2, 1/2) \in \mathbb{R}^2$, fair coin

$p = (1/3, 2/3) \in \mathbb{R}^2$, unfair coin

Empirical distribution

- Estimate of the probability distribution based on samples of the distribution
  - Given samples $\{s^i\}_{i=1}^N, s^i \in S$, the empirical distribution is $\hat{p}(s) = \frac{1}{N} \sum_s 1_{\{s=s^i\}}$
  - Recall the indicator function is

$$1_{\{s=s^i\}} = \begin{cases} 1 & \text{if } s = s^i, \\ 0 & \text{otherwise .} \end{cases}$$

independently drawn from the distribution

- Example: flip a coin 5 times $S = \{s_1 = \text{heads}, s_2 = \text{tails}\}$

$\{s^1 = s_1, s^2 = s_1, s^3 = s_2, s^4 = s_1, s^5 = s_2\}$

$\hat{p}(s) = \frac{1}{5} \sum_{i=1}^{5} 1_{\{s=s^i\}} = \begin{cases} s = \text{heads} & 3/5 \\ s = \text{tails} & 2/5 \end{cases}$

Probabilistic interpretation of data

- Consider the data $\{x^i, y^i\}_{i=1}^N$, $x^i \in X, y^i \in Y$. We can think of data coming from an unknown probability distribution $D(X \times Y)$

regression $\quad X = \mathbb{R}^d, \quad Y = \mathbb{R}^m$

classification $\quad X = \mathbb{R}^d, \quad Y = \{1, 2, \cdots, K\}$

- Conditional distribution of labels given a feature, $p(y \mid x)$, is a distribution on Y given x

  - Example: in our logistic regression, $z = b + w_1 x_1 + \cdots + w_d x_d$

$p(y=0 \mid x)$, $p(y=1 \mid x)$ — true probabilities

$\hat{p}(y=0 \mid x) = 1 - \sigma(z)$, $\hat{p}(y=1 \mid x) = \sigma(z)$ — estimate for p, given logistic classifier

Empirical expectation

- Given a distribution $p(s)$, the expectation of a function $f : S \to \mathbb{R}$ is $\mathbb{E}_p[f] = \sum_s p(s)f(s)$

continuous domain S, $\mathbb{E}_p[f] = \int_S f(s)\,p(s)\,ds$.

p: density

- The empirical expectation given samples $\{s^i\}_{i=1}^N$ is

$$\sum_x \hat{p}(s)f(s) = \frac{1}{N}\sum_s 1_{\{s=s^i\}} f(s) = \frac{1}{N} f(s^i) = \frac{1}{N}\sum_{i=1}^{N} f(s^i)$$

- Example: in our linear regression, we have been minimizing the empirical loss

$\frac{1}{N}\sum_{i=1}^{N}(y^i - \hat{y}^i)^2$ instead of the true loss $\mathbb{E}_D(y - \hat{y})^2$ (since true loss is unknown)

Cross entropy of a distribution relative to another

- Cross entropy: given two distributions $p, q : \{1,2,\ldots,K\} \to \mathbb{R}$ is $\mathbb{E}_p[\log(q)]$

$$H(p,q) = -\sum_{j=1}^{K} p(j) \log(q(j))$$

used as measure of distance between distributions.

- In logistic regression, our two distributions are the true and the predicted distributions of the labels given features: $p(y \mid x), \hat{p}(y \mid x)$

$p(y=0|x)$

$p(y=1|x)$

our prediction

Cross entropy example and interpretation

logische Loss

$$\hat{L}(w,b) = -\frac{1}{N}\sum_{i=1}^{N}\left(y^i \log \hat{p}^i + (1-y^i)\log(1-\hat{p}^i)\right), \quad \hat{p}^i = \sigma(z^i),\ z^i = b + w_1 x_1^i + \ldots + w_d x_d^i$$

$\hat{p}^i$: probability of label 1

$1-\hat{p}^i$: probability of label 0

- Example

$$y = 1, \hat{p} = 0.9 \Rightarrow -\log(0.9) \approx 0.10$$
$$y = 1, \hat{p} = 0.1 \Rightarrow -\log(0.1) \approx 2.30$$
$$y = 0, \hat{p} = 0.1 \Rightarrow -\log(1-0.1) \approx 0.10$$
$$y = 0, \hat{p} = 0.9 \Rightarrow -\log(1-0.9) \approx 2.30$$

- Cross-entropy is small when the model assigns high probability to the correct label, and large otherwise.

⇒ it's a good idea to minimize cross-entropy between our prediction distribution, and the true one.

Logistic loss
Interpretation as binary cross-entropy loss

- By interpreting the sigmoid function as probability of class 1, $\sigma(z) = \frac{1}{1+e^{-z}}$.

prob. of class 0, $1 - \sigma(z) = \frac{e^{-z}}{1+e^{-z}}$

- the *empirical* cross-entropy between the predicted distribution and true distribution becomes

$$-\frac{1}{N}\sum_{i=1}^{N}\left[(1-y^i)\log\frac{e^{-z^i}}{1+e^{-z^i}} + y^i \log\frac{1}{1+e^{-z^i}}\right]$$

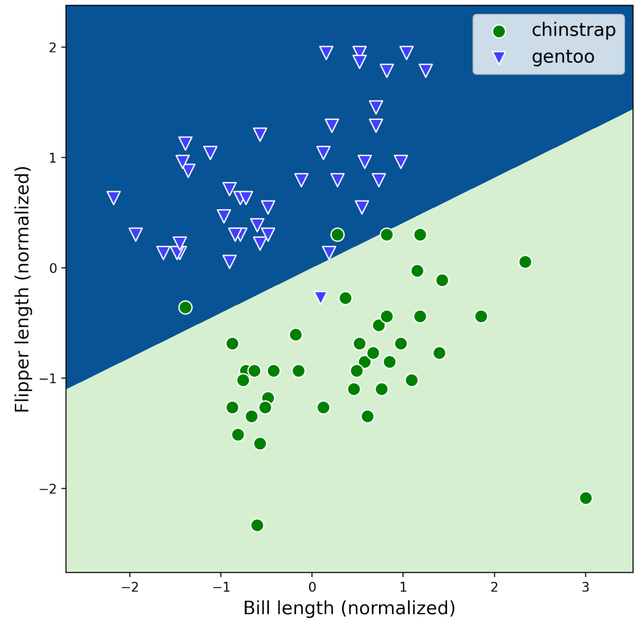
- Verify that this is precisely the logistic loss function.

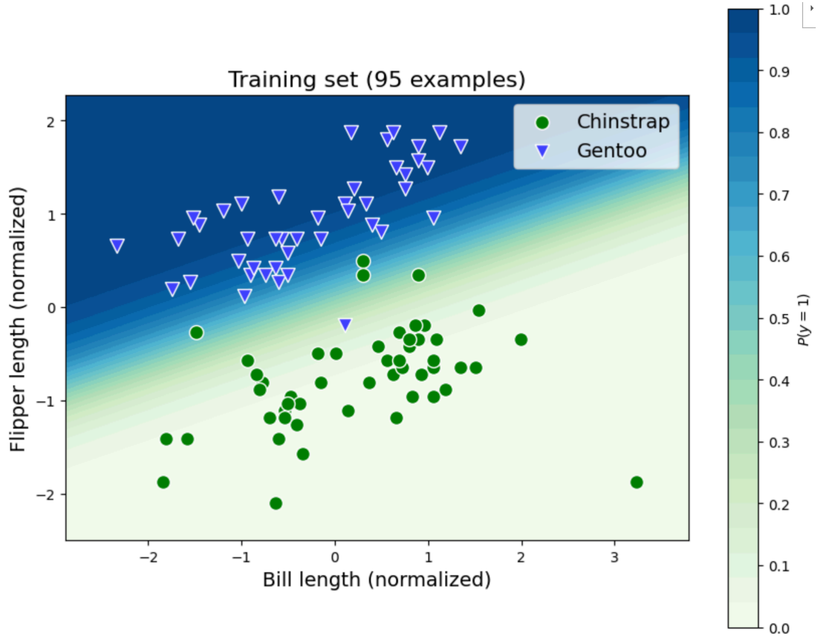
check with equation *** on slide 12.

Logistic regression output

Palmer Penguins

	species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
0	Chinstrap	49.0	19.5	210.0	3950.0
1	Chinstrap	50.9	19.1	196.0	3550.0
2	Gentoo	42.7	13.7	208.0	3950.0
3	Chinstrap	43.5	18.1	202.0	3400.0
4	Chinstrap	49.8	17.3	198.0	3675.0





Based on dataset available here: https://www.kaggle.com/code/parulpandey/penguin-dataset-the-new-iris

Performance metrics for binary classification

- True label versus predicted label

$y = 0, \quad \hat{y} = 0,$ true negative, number of true negatives C_{tn}

$y = 0, \quad \hat{y} = 1,$ false positive, number of false positives C_{fp}

$y = 1, \quad \hat{y} = 0,$ false negative, number of false negatives C_{fn}

$y = 1, \quad \hat{y} = 1,$ true positive, number of true positives C_{tp}

N data point

prediction

- Accuracy $\frac{C_{tn} + C_{tp}}{N}$

- Error rate $\frac{C_{fn} + C_{fp}}{N}$

- Recall $\frac{C_{tp}}{C_{tp} + C_{fn}}$ how many of true positives we caught.

- Confusion matrix

true label

C_{tp}	C_{fn}
C_{fp}	C_{tn}

Exercise - logistic regression

- You test a classifier on 100 machines. From these, 20 machines are faulty and 80 are healthy. Let us use “positive” for a faulty machine and “negative” for a healthy machine. The classifier outputs: 18 true positives (catches 18 of the 20 faults) and 5 false positives.
- Write the confusion matrix.
- Determine the accuracy, error rate and recall.

predicted

true label		
	18	2
	5	75

Note ... the entries add up to 100

accuracy: $\frac{18+75}{100} = 93\%$

error rate: $1 - 93\% = 7\%$

recall: $\frac{18}{20} = 90\%$

Exercise (check in exercise hour)
Performance metric for binary classification

- We have used two approaches to train classifiers for spam email detection: “non-spam” (class 0) and “spam” (class 1).
- Our test set has1000 emails, 900 of which were non-spam.
- Approach 1: classified all data as non-spam.
- Approach 2: classified 850 of non-spam emails as non-spam and 50 of spam emails as spam.

1. Write the confusion matrix of each approach.
2. Compute the error rate and accuracy of each algorithm.
3. Which classifier is better from your perspective?

Solution

- Approach 1
- Confusion matrix

- Accuracy

- Error rate

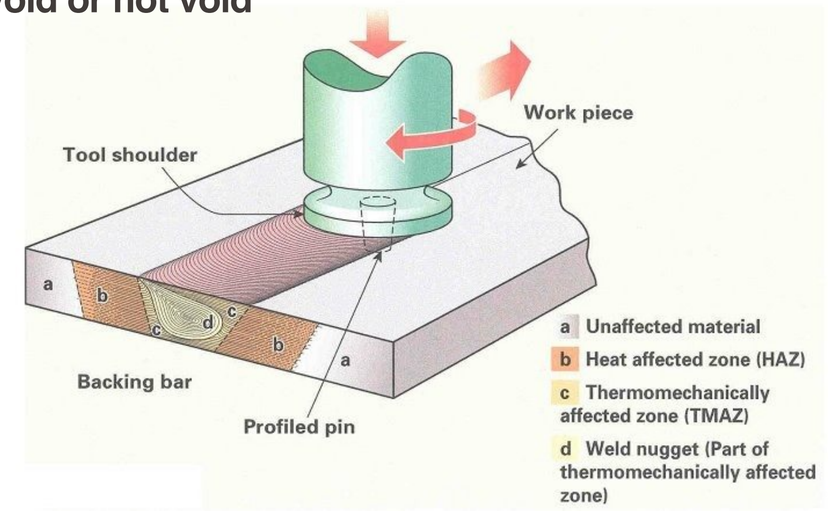
- Approach 2
- Confusion matrix

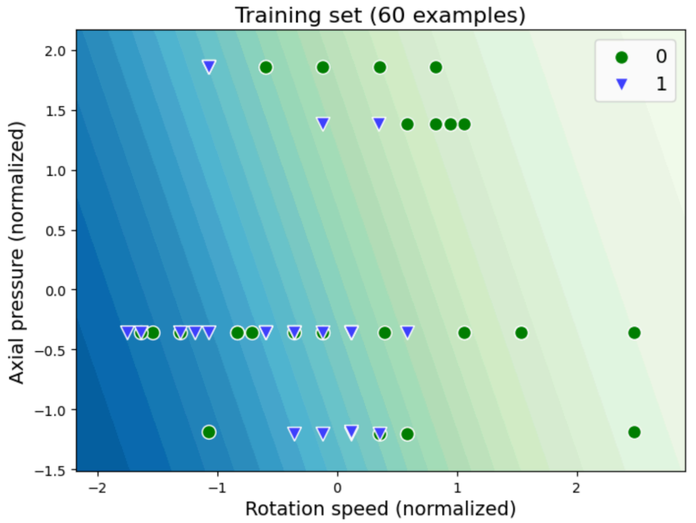
- Accuracy

- Error rate

Logistic regression in python

- Dataset 1: Void Formation in Welding, based on the paper
- Goal: formation of voids in friction stir welding as a function of the operation conditions
  - Tool rotational speed, axial pressure
  - The label: void or not void



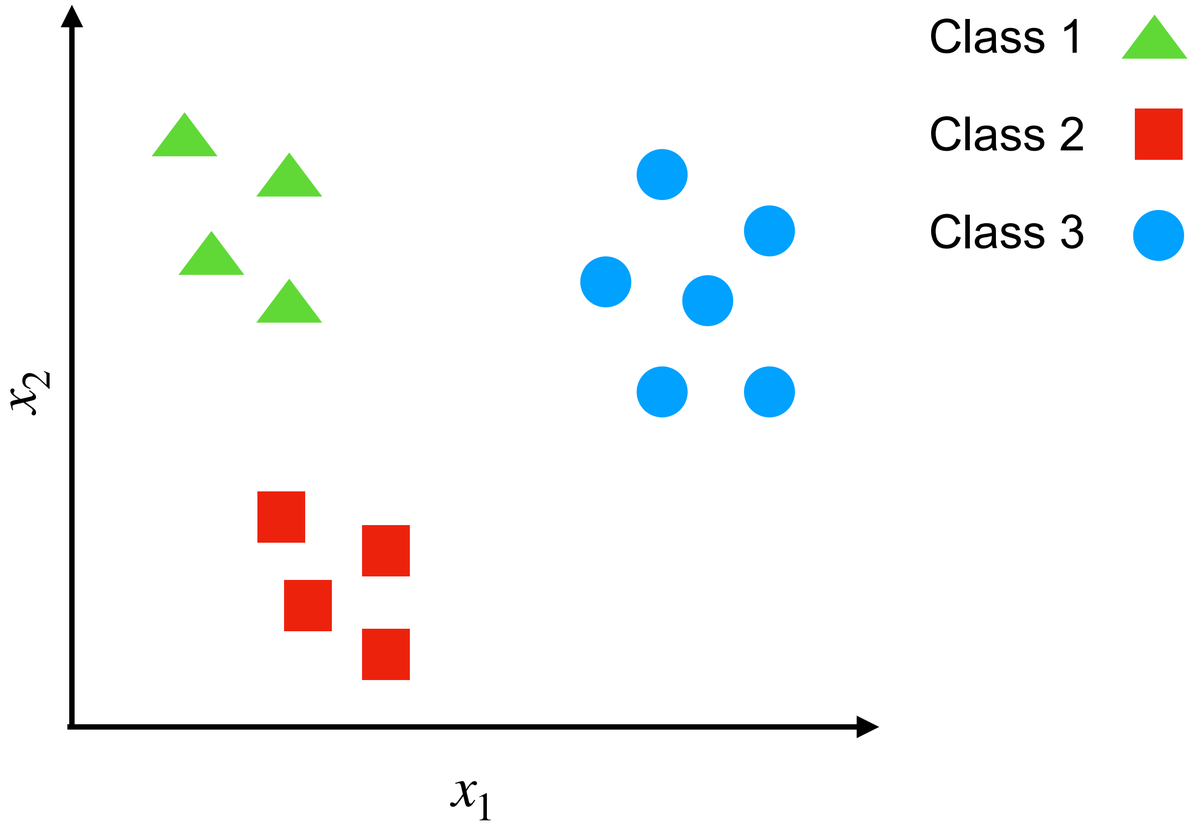


- Dataset 2: discriminate between sonar signals bounced off a mine (metal cylinder) and those bounced off a roughly cylindrical rock
- Goal: predict whether the object is mine or rock based on
  - The features (60 of them) are the energy within a particular frequency band, integrated over a certain period of time
  - The label: rock/mine

Multinomial logistic regression predictor

Example: Medical diagrams:

Not ill (y = 1),Cold(y = 2), Covid(y = 3)



1) Associate a line (hyperplane in higher dimension) for each class:

$$z_1 = w_{1,1}x_1 + w_{1,2}x_2 + b_1$$

$$z_2 = w_{2,1}x_1 + w_{2,2}x_2 + b_2$$

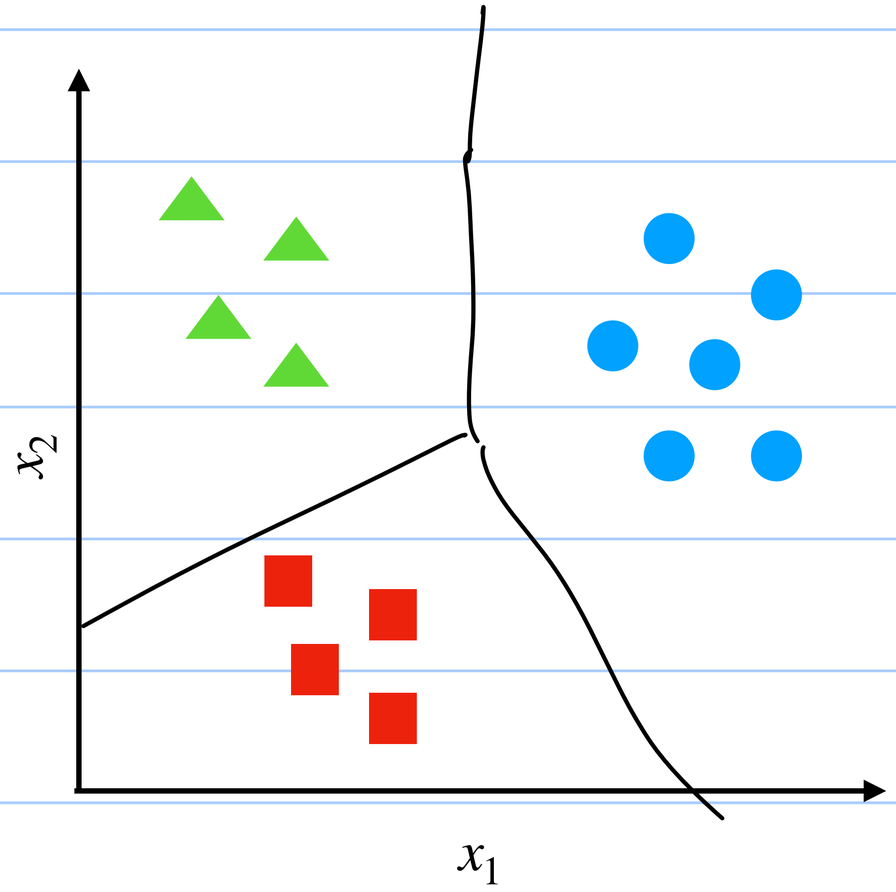
$$z_3 = w_{3,1}x_1 + w_{3,2}x_2 + b_3$$

2) For a given $x^i \in \mathbb{R}^2$, Predict the class with the largest value

$$\hat{y}^i = \arg \max_{k \in \{1,2,3\}} w_{k,1}x_1^i + w_{k,2}x_2^i + b_k$$

Multinomial logistic regression decision boundaries

$$\hat{y}^i = \arg \max_{k \in \{1,2,3\}} w_{k,1} x_1^i + w_{k,2} x_2^i + b_k$$



decision boundary, between class k, l:

$$w_{l,1} x_1 + w_{l,2} x_2 + b_l =$$

$$w_{k,1} x_1 + w_{k,2} x_2 + b_k$$

$$(w_{l,1} - w_{k,1}) x_1 + (w_{l,2} - w_{k,2}) x_2 +$$

$$b_l - b_k = 0$$

Probabilistic prediction

Extend the logistic function to K-class setting by defining the softmax function

For each class $c \in \{1,2,\ldots,K\}$, define $z_c = w_{c,1}x_1 + \ldots + w_{c,d}x_d + b_c$

$\{x^i, y^i\}$

$x^i \in \mathbb{R}^d$, $y^i \in \{1,2,\cdots,K\}$

$$softmax : \mathbb{R}^K \to (0,1)^K, \quad softmax(z) = \left(\frac{\exp(z_1)}{\sum_{j=1}^K \exp(z_j)}, \ldots, \frac{\exp(z_K)}{\sum_{j=1}^K \exp(z_j)}\right)$$

$\exp(z) = e^z$

Example: $softmax([1,5,2,3]) = [0.0152,0.8310,0.0414,0.1125]$

$e^z \geq 0$, $\sum_{j=1}^K e^{z_j} \geq 0 \Rightarrow \frac{e^{z_j}}{\sum_{j=1}^K e^{z_j}} \geq 0$

$\sum_{c=1}^K \frac{e^{z_c}}{\sum_{j=1}^K e^{z_j}} = \frac{\sum_c e^{z_c}}{\sum_{j=1}^K e^{z_j}} = 1$

Training multinomial logistic regression

- Multinomial cross-entropy loss for multiclass ($K > 2$) classification

$$J(\mathbf{w}, b) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} \mathbf{1}_{\{y^{(i)}=c\}} \log\left(\frac{\exp(z_c^i)}{\sum_{j=1}^{K}\exp(z_j^i)}\right)$$

$\{x^i, y^i\}$, $x^i \in \mathbb{R}^d$, $y^i \in \{1, 2, \dots, K\}$

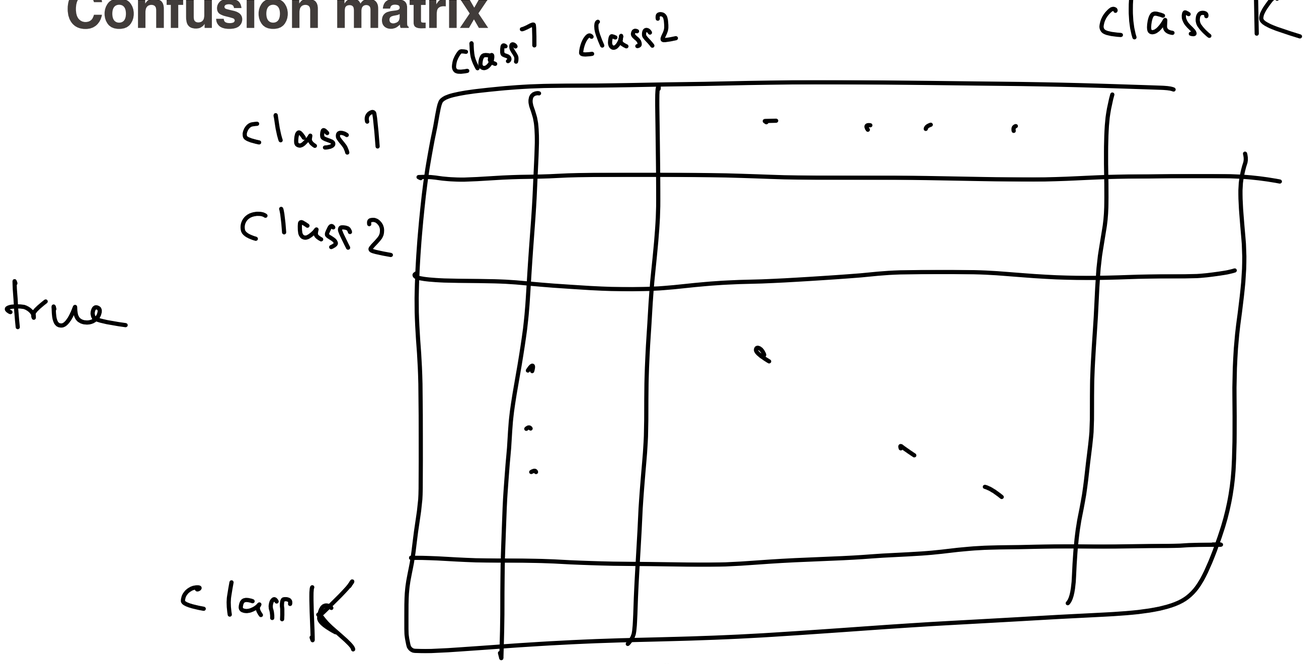
$$z_c^i = b_c + w_{c,1} x_1^i + \dots + w_{c,d} x_d^i$$

- Verify that the loss is differentiable and convex.

we can use gradient descent to find parameters that minimize the loss $\{(b_c, w_{c,1}, \dots, w_{c,d})\}_{c \in \{1,2,\dots,K\}}$

Performance metric

Confusion matrix

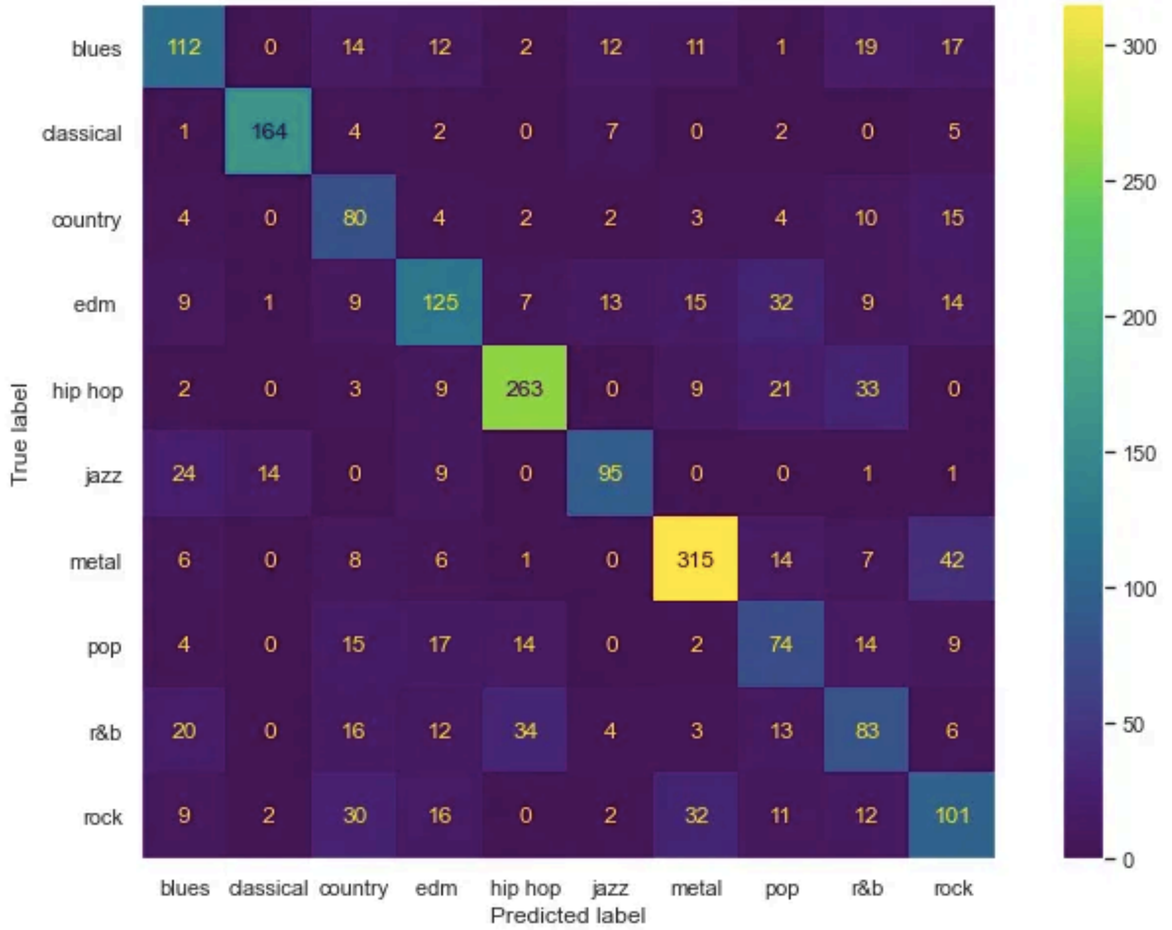


Accuracy

$$\text{Accuracy} = \frac{\text{sum of diagonals}}{\#\text{ of test set}}$$

Error rate

$$1 - \text{Accuracy}$$



Example from genre classification based on music data

Exercise

- The goal is to determine reason for faults in a machine, whether they were due to mechanical, electrical, software failures. Given 150 datapoints, 90 were used for training, 30 for validation, and 30 for testing. Based on the training and validation, a classifier was developed. The performance compared to ground truth is given as follows.

test set performance

Confusion matrix

		Electrical	Mechanical	Software
	Electrical	8	1	1
	Mechanical	2	7	1
	Software	1	2	7

10 mechanicals

- Determine the accuracy.

22/30

- What's the percentage of mechanical faults correctly identified?

7/10 = 70%

Summary

- Logistic regression for classification
  - Binary classification, logistic loss
  - Multinomial classification
  - Cross entropy interpretation
  - Accuracy, error rate, recall

- Probability
  - Probability distribution, expectation
  - Empirical distribution

Your tasks this week

- Go through lecture 3 slides
- Do the exercises in lecture slides
- Do the python exercises posted on Moodle
- Bring your questions to exercise hour