

Mudcard

- With so many different options for regression and classification metrics, how will we know which one to pick? Or should we be trying many/all metrics in each analysis?
- Which regression metrics should we choose when facing a specific case? or we just can choose any one of these?
- How do you know when to use which evaluation metric? For example, how do you choose between MSE, RMSE, MAE, and R2 score?
- I am curious how we should best choose which loss function to use since there are so many ways to evaluate it.
 - careful, the loss function is something else we will cover today
 - yes, there are a lot of evaluation metrics especially for classification
 - you need to consider what you want to use your ML model for, that usually determines which metric (or metrics) are the best
 - you can sometimes exclude metrics based on the target variable (e.g., do not use accuracy or the ROC if you have an imbalanced classification problem)
 - in regression, it usually matters less which metric you chose, it's usually up to your subjective preference
 - personally I often use RMSE because it has the same unit as the target variable, or R2 because it is dimensionless
- In the extreme cases section for logloss, why are we subtracting the log term by 1 for class-0 points in the perfect classifier example? The equation I see is $(1 - 0) * (1 - \ln(1 - 10^{-15}))$. I am not sure where that extra '1 -' is coming from.
 - for class 0 points, $y_{\text{true}} = 0$, so the second term in the logloss equation remains, that's where the extra '1 -' comes from
- why we set p_{crit} to the sorted prediction? I thought it should be 0.5!"
 - 0.5 is the default in sklearn when you need the predicted classes
 - but you need to iterate through various p_{crit} values to determine the ROC and P-R curves
 - one p_{crit} only gives one point on the curves
- Can we use AUC-ROC/PR curves for a multi-class problem?
 - yes but it is not straightforward
 - read more [here](#)
- Why does unbalanced data make ROC bad? Namely, why does using TN give worse results?
 - it's the same argument we had for accuracy
 - the baseline ROC AUC is already pretty close to a perfect classifier so there isn't much room for improvement
 - this is generally true for any metric that contains TN

- TN is the true negatives, the number of correctly classified points in the majority class, which is large even for a baseline classifier
- **Is there a visualization you could show of the different kinds of possible train and validation curves based on hyper-parameter tuning we might see?**
 - yes, you will see several such curves during the next couple of weeks and you will create plots in the problem sets as well
- **Do you use all of your training samples (or only some) when calculating RMSE on your trained regression model?**
 - all
- **what is the easiest way to compute TP TN FP and FN when we need a confusion matrix in order to produce fpr and R**
 - sklearn's confusion_matrix function
- **How to handle text variables in preprocessing?**
 - OneHotEncoder or the OrdinalEncoder
 - please rewatch the week 4 lecture on preprocessing, we had a whole lecture on this
- **The muddiest part was understanding the meaning of the logloss output. Is there a way to aptly describe the metric in a business sense?**
 - logloss measures how close the predicted probabilities are to the true labels
 - the more the predicted probabilities differ from the true values, the higher the logloss metric is
 - I think this is relatively straightforward
 - what I struggle with is explaining the f score to business people
- **When should we use logloss metric?**
 - when you have extremely limited resources to act and you want to make sure you allocate the resources to the people who need it most
 - in any scenario when you care about accurately predicting the probabilities
- ****I am confused about the difference between the ROC curve and the precision-recall curve.**
 - read more [here](#) and [here](#)
- **Does the model perform better when the ROC curves closer towards the upper left corner?**
 - depends on your definition of 'better' :)
- **Are we assuming that all $- \log(\text{pred_prob_class1})$ will be smaller than 1?**
 - no, sometimes it can be 1 if the model is confident
- **Does F-1 score and other metrics generalise to more than 2 class classification problems?**
 - you will see in PS6 :)
- **What exactly are class 0 and class 1?**
 - class 0 is usually the majority class and class 1 is the minority class
 - is the patient sick? if the answer is 'no', that's class 0. if the answer is 'yes', that's class 1.

- will the customer buy the product? if the answer is 'no', that's class 0. if the answer is 'yes', that's class 1.
- will the reader click on the ad? if the answer is 'no', that's class 0. if the answer is 'yes', that's class 1.
- is the transaction fraudulent? if the answer is 'no', that's class 0. if the answer is 'yes', that's class 1.
- in your classification dataset, the target variable is 0s and 1s

Supervised ML algorithms, part 1

By the end of this lecture, you will be able to

- describe the main components of any ML algorithm
- describe how linear regression works
- describe how logistic regression works

Supervised ML algorithms, part 1

By the end of this lecture, you will be able to

- **describe the main components of any ML algorithm**
- describe how linear regression works
- describe how logistic regression works

The supervised ML pipeline

The goal: Use the training data (X and y) to develop a **model** which can **accurately** predict the target variable (y_new) for previously unseen data (X_new).

1. Exploratory Data Analysis (EDA): you need to understand your data and verify that it doesn't contain errors

- do as much EDA as you can!

2. Split the data into different sets: most often the sets are train, validation, and test (or holdout)

- practitioners often make errors in this step!
- you can split the data randomly, based on groups, based on time, or any other non-standard way if necessary to answer your ML question

3. Preprocess the data: ML models only work if X and Y are numbers! Some ML models additionally require each feature to have 0 mean and 1 standard deviation (standardized features)

- often the original features you get contain strings (for example a gender feature would contain 'male', 'female', 'non-binary', 'unknown') which needs to be transformed into numbers
- often the features are not standardized (e.g., age is between 0 and 100) but it needs to be standardized

4. Choose an evaluation metric: depends on the priorities of the stakeholders

- often requires quite a bit of thinking and ethical considerations

5. Choose one or more ML techniques: it is highly recommended that you try multiple models

- start with simple models like linear or logistic regression
- try also more complex models like nearest neighbors, support vector machines, random forest, etc.

6. Tune the hyperparameters of your ML models (aka cross-validation)

- ML techniques have hyperparameters that you need to optimize to achieve best performance
- for each ML model, decide which parameters to tune and what values to try
- loop through each parameter combination
 - train one model for each parameter combination
 - evaluate how well the model performs on the validation set
- take the parameter combo that gives the best validation score
- evaluate that model on the test set to report how well the model is expected to perform on previously unseen data

7. Interpret your model: black boxes are often not useful

- check if your model uses features that make sense (excellent tool for debugging)
- often model predictions are not enough, you need to be able to explain how the model arrived to a particular prediction (e.g., in health care)

Supervised ML algorithms: three parts

- 1. a **mathematical model** (f) is used to convert the feature values into a prediction

$f(X_i) = y'_i$, where i is the i th data point in our sample. X_i is a vector and y'_i is a number. - f is your supervised ML algorithm - it usually has a number of intrinsic parameters

- 2. an **optimization algorithm** is used to determine the intrinsic parameter values given the training set
 - there are various algorithms
 - e.g., analytical solutions, brute force, gradient descent, backpropagation
- 3. the optimization algorithm minimizes a metric called **the cost function**

- the cost function is used to determine the best intrinsic parameters of one model based on the training data
- it is not the same as the evaluation metric
 - you use the evaluation metric to compare various models
 - the model uses the cost function to find the best values of its intrinsic parameters
 - keep in mind though that the same metric can be used as the cost function and the evaluation metric (e.g., MSE in regression) but that's not necessarily the case (e.g., the cost function is MSE but you use R2 as an evaluation metric).

Supervised ML algorithms, part 1

By the end of this lecture, you will be able to

- describe the main components of any ML algorithm
- **describe how linear regression works**
- describe how logistic regression works

Linear Regression

```
In [1]: from sklearn.linear_model import LinearRegression # import the model
LinReg = LinearRegression() # initialize a simple linear regression model
LinReg.fit(X_train,y_train) # we will learn now what happens when you issue this
```

- This is the mathematical model:

$$\begin{aligned} f(X_i) = y'_i &= w_0 + X_{i1}w_1 + X_{i2}w_2 + \dots \\ &= w_0 + \sum_{j=1}^m w_j X_{ij} \end{aligned}$$

where y'_i is the prediction of the linear regression model and w are parameters.

- The cost function is MSE - We will find the best parameter values by brute force first, then simple gradient descent.

Let's generate some data

```
In [2]: # load packages and generate data
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import matplotlib
from sklearn.datasets import make_regression
matplotlib.rcParams.update({'font.size': 13})

# fix the seed so code is reproducible
np.random.seed(1)
```

```
# generate n_samples points
n_samples = 30

# generate data
X, y, coef = make_regression(n_samples = n_samples, n_features = 1, coef = True)
print(coef) # the coefficients of the underlying linear model, the bias is 0.
print(np.dot(X,coef)[: ,0]) # noise is added to the label
print(y)
df = pd.DataFrame()
df['x1'] = X[:,0]
df['y'] = y
print(df.head())
df.to_csv('data/regression_example.csv',index=False)
```

```
28.777533858634875
[  9.18115839 -30.87739085  24.90429735 -4.96205858  32.94232532
  14.46054785 -66.23260778 -19.67600162 -11.0521372 -25.26260034
 -59.28576902 -31.67310564 -31.65215819  32.62708851  50.21137962
 -3.53647763 -26.92913658  46.7446537  42.07586066  25.94555749
 -21.90565736  15.2623224 -15.19948048 -7.70915828  16.77198455
   1.21480753  25.92441258 -7.17626442 -17.60484091 -9.27837201]
[  3.08130585 -48.2639338  25.93592732  1.87178372  45.4428258
   6.88380644 -72.44927625 -16.20740827 -12.15967691 -25.0337406
 -66.76530644 -35.83422144 -27.17695128  53.04737598  52.96856006
  10.13384942 -20.19197583  36.16208808  58.99412194  38.42855536
 -32.81240623  6.77988325 -15.82439642 -20.62478531  18.61129032
   9.78715027  31.80735422 -4.11214063 -11.04200012 -15.08634424]
      x1      y
0  0.319039  3.081306
1 -1.072969 -48.263934
2  0.865408  25.935927
3 -0.172428  1.871784
4  1.144724  45.442826
```

In [3]:

```
def predict(X,w):
    if len(np.shape(w)) != 2:
        w = np.array(w)[np.newaxis,:] # just a numpy trick to make the dot product work
    y_pred = w[0,0] + X.dot(w[0,1:]) # intercept + w_i*x_i
    return y_pred

def cost_function(X,y_true,w):
    """
    Take in a numpy array X,y_true, w and generate the cost function
    of using w as parameter in a linear regression model
    """
    m = len(y)
    w = np.array(w)[np.newaxis,:] # just a numpy trick to make the dot product work
    y_pred = predict(X,w)
    cost = (1/m) * np.sum(np.square(y_true-y_pred)) # this is MSE
    return cost
```

$$y'_i = w_0 + x_{i1}w_1$$

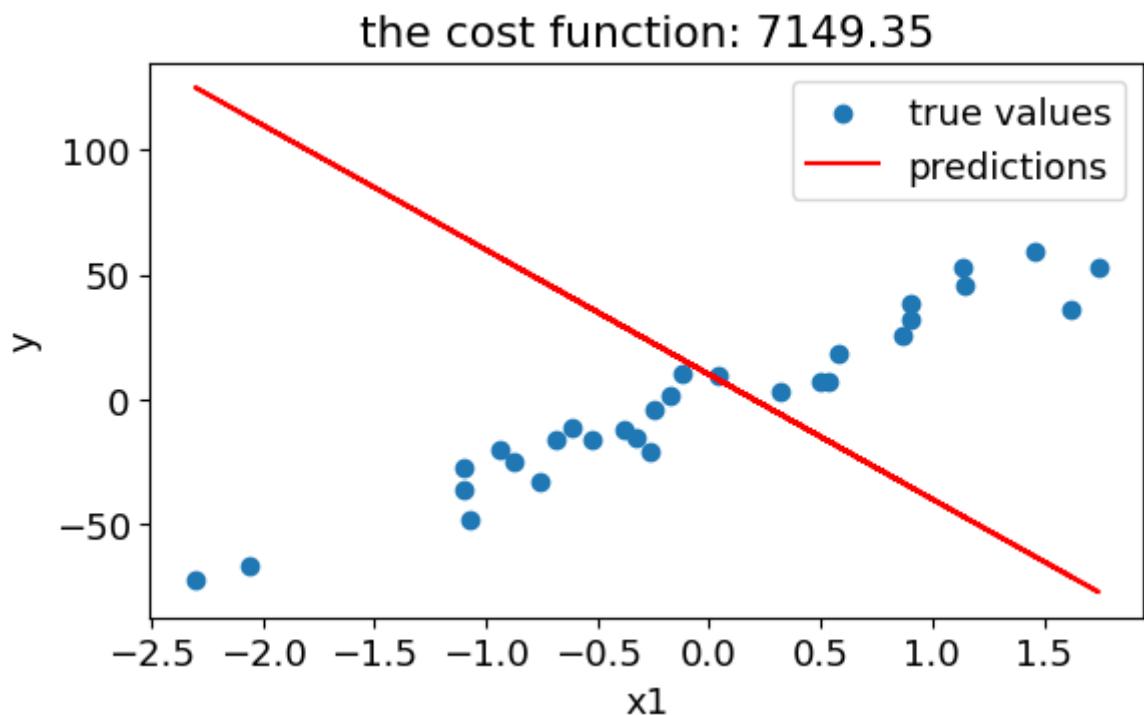
- w_0 is the intercept
- w_1 is the slope

We are looking for the best fit line!

For a given w vector, the cost function returns the MSE.

```
In [4]: w = [10,-50] # intercept is w[0], the slope is w[1]

plt.figure(figsize=(6.4,3.6))
plt.scatter(df['x1'],df['y'],label='true values')
plt.plot(df['x1'],predict(df['x1'].values[:,np.newaxis],w),label='predictions',
plt.title('the cost function: '+str(np.around(cost_function(df['x1'].values[:,r
plt.xlabel('x1')
plt.ylabel('y')
plt.legend()
plt.savefig('figures/line_fit.png',dpi=300)
plt.show()
```



What we want:

- Find the w vector that minimizes the cost function!
 - that's our best fit model

How we do it:

- brute force
 - create a grid of $w[0]$ and $w[1]$ values
 - loop through all w vectors on the grid
 - find the w vector that comes with the smallest cost

```
In [5]: n_vals = 101
```

```

w0 = np.linspace(-100,100,n_vals) # the intercept values to explore
w1 = np.linspace(-100,100,n_vals) # the slope values to explore

cost = np.zeros([len(w0),len(w1)]) # the cost function's value for each w

# loop through all intercept-slope combinations and calculate the cost function
for i in range(n_vals):
    for j in range(n_vals):
        w = [w0[i],w1[j]]
        cost[i,j] = cost_function(df['x1'].values[:,np.newaxis],df['y'],w)

print('min(cost):',np.min(cost))
min_coords = np.unravel_index(cost.argmin(),np.shape(cost))
print('best intercept:',w0[min_coords[0]])
print('best slope:',w1[min_coords[1]])

```

```

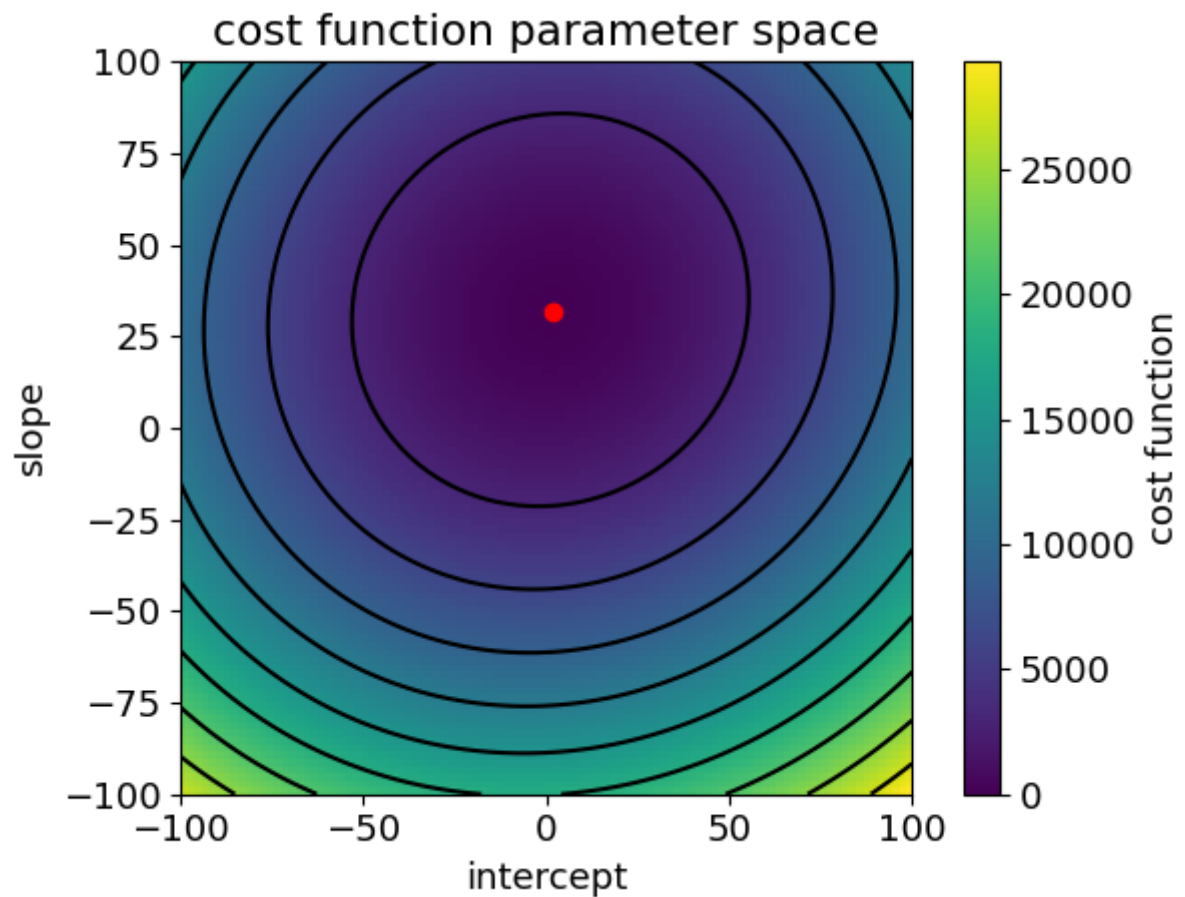
min(cost): 71.43643291686587
best intercept: 2.0
best slope: 32.0

```

```

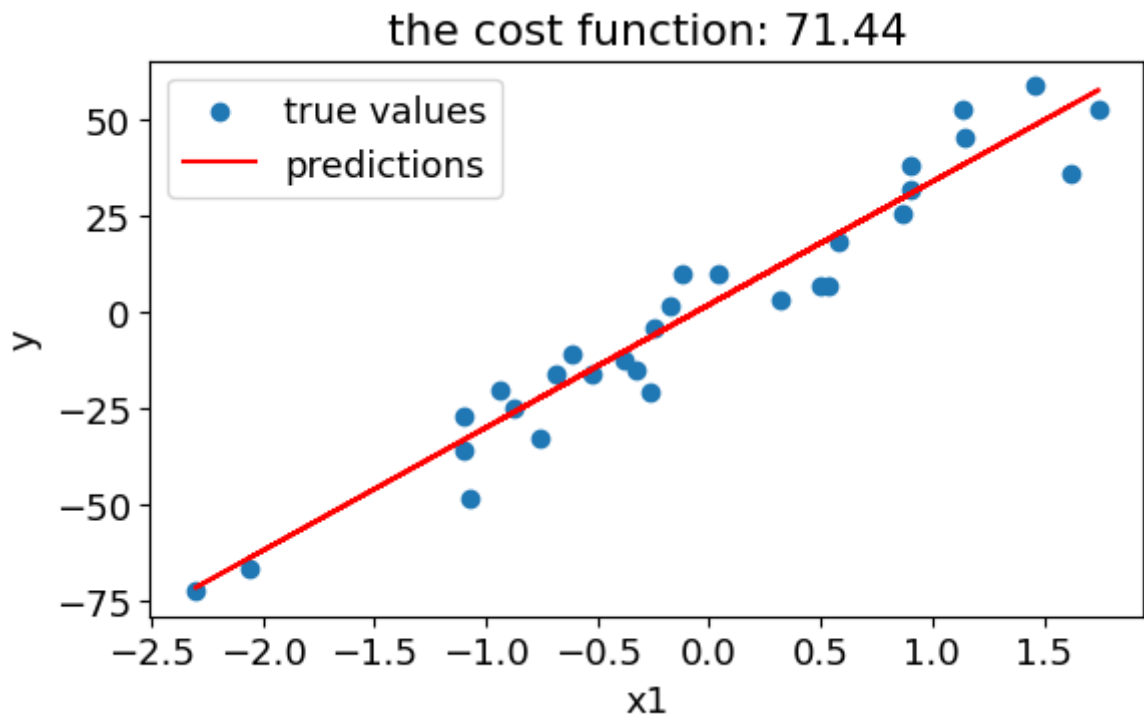
In [6]: plt.figure(figsize=(6.4,4.8))
ax = plt.gca()
extent = (np.min(w0),np.max(w0),np.min(w1),np.max(w1))
fig = ax.imshow(cost.T,origin='lower',extent=extent,vmin=0)
plt.colorbar(fig,label='cost function')
ax.contour(w0,w1,cost.T,levels=10,colors='black')
plt.scatter(w0[min_coords[0]],w1[min_coords[1]],c='r')
ax.xaxis.set_ticks_position("bottom")
plt.xlabel('intercept')
plt.ylabel('slope')
plt.title('cost function parameter space')
plt.tight_layout()
plt.savefig('figures/cost_function.png',dpi=300)
plt.show()

```

```
In [7]: w = [2,32] # intercept is w[0], the slope is w[1]

plt.figure(figsize=(6.4,3.6))
plt.scatter(df['x1'],df['y'],label='true values')
plt.plot(df['x1'],predict(df['x1'].values[:,np.newaxis],w),label='predictions',
plt.title('the cost function: '+str(np.around(cost_function(df['x1'].values[:,r
plt.xlabel('x1')
plt.ylabel('y')
plt.legend()
plt.savefig('figures/line_fit.png',dpi=300)
plt.show()
```



Quiz 1

The brute force approach works but...

- the number of w vectors to loop through explodes with the number of features we have
 - with n features, we would need to loop through $\sim 100^n$ w vectors.
 - no guarantee that the best w vector is within our grid.
- We need to use a smarter numerical method to find the best w !
 - gradient descent to the rescue!

How do we find the best w values?

- start with arbitrary initial w values and the cost function L
- repeat until convergence:

$$w_j := w_j - l \frac{\partial L(w)}{\partial w_j},$$

where $\frac{\partial L(w)}{\partial w_j}$ is the gradient of the cost function at the current w location and l is the learning rate.

- the gradient tells us which way the cost function is the steepest
- the learning rate tells us how big of a step we take in that direction

```
In [8]: def gradient_descent(X,y_true,w, learning_rate=0.01, iterations=100):
        ...
        X    = Matrix of X
        y    = Vector of Y
```

```

w = Vector of ws
learning_rate
iterations = no of iterations

Returns the final w vector and array of cost history over no of iterations
'''
m = len(y_true)
w = np.array(w)[np.newaxis,:]

cost_history = np.zeros(iterations)
w_history = np.zeros([iterations,np.shape(w)[1]])
for it in range(iterations):

    y_pred = predict(X,w)
    delta_w = np.zeros(np.shape(w)) # the step we take
    # the derivative of the cost function with respect to the intercept
    delta_w[0,0] = (1/m) * sum(y_pred - y_true) * learning_rate
    # the derivative of the cost function with respect to the slopes * learning_rate
    delta_w[0,1:] = (1/m)*learning_rate*( X.T.dot((y_pred - y_true)))
    w = w - delta_w # update w so we move down the gradient
    w_history[it] = w[0]
    cost_history[it] = cost_function(X,y_true,w[0])

return w[0], cost_history, w_history

```

```

In [9]: w,cost_history,w_hist = gradient_descent(df['x1'].values[:,np.newaxis],df['y'],
print(w)
print(w_hist)

```

```

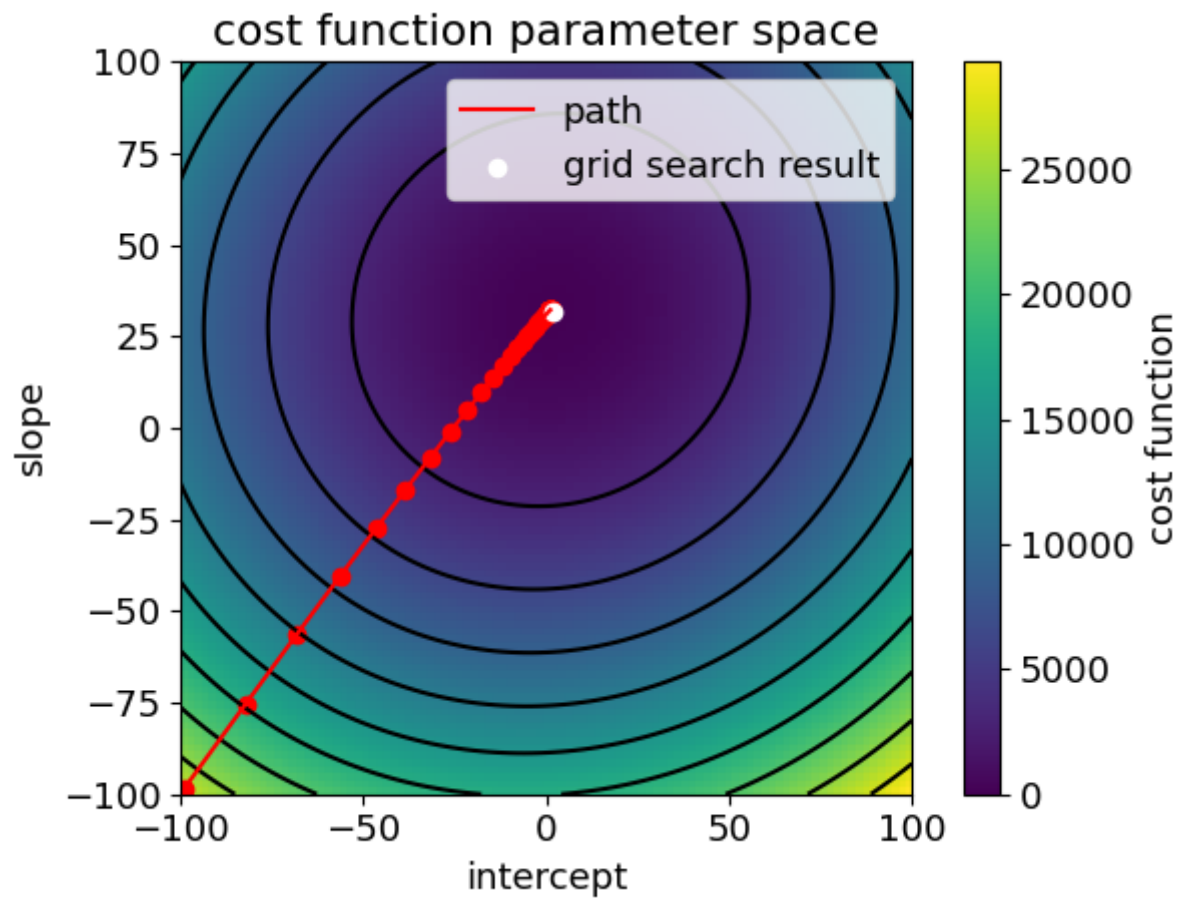
[ 1.14458074 32.24849231]
[[-99.06782181 -98.70917181]
 [-98.14419169 -97.43097591]
 [-97.22903173 -96.16528833]
 ...
 [ 1.14457933 32.24849104]
 [ 1.14458004 32.24849168]
 [ 1.14458074 32.24849231]]

```

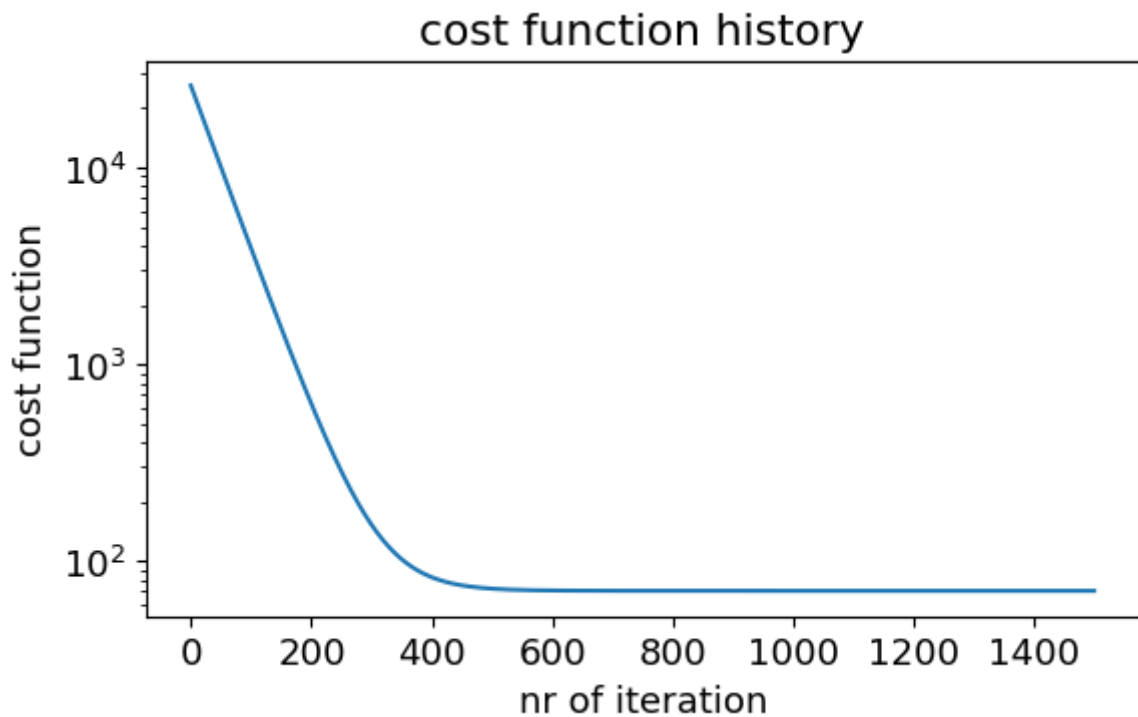
```

In [10]: plt.figure(figsize=(6.4,4.8))
ax = plt.gca()
extent = (np.min(w0),np.max(w0),np.min(w1),np.max(w1))
fig = ax.imshow(cost.T,origin='lower',extent=extent,vmin=0)
plt.colorbar(fig,label='cost function')
ax.contour(w0,w1,cost.T,levels=10,colors='black')
plt.plot(w_hist[:,20,0],w_hist[:,20,1],color='r',label='path')
plt.scatter(w_hist[:,20,0],w_hist[:,20,1],c='r')
plt.scatter(w0[min_coords[0]],w1[min_coords[1]],c='w',label='grid search result')
ax.xaxis.set_ticks_position("bottom")
plt.legend()
plt.xlabel('intercept')
plt.ylabel('slope')
plt.title('cost function parameter space')
plt.tight_layout()
plt.savefig('figures/cost_function_with_path.png',dpi=300)
plt.show()

```



```
In [11]: plt.figure(figsize=(6.4,3.6))
plt.plot(cost_history)
plt.semilogy()
plt.ylabel('cost function')
plt.xlabel('nr of iteration')
plt.title('cost function history')
plt.savefig('figures/cost_hist.png',dpi=300)
plt.show()
```



DO NOT USE MY `gradient_descent` FUNCTION!

- it is for illustration purposes only
- it is much slower than the sklearn implementation!

Quiz 2

Supervised ML algorithms, part 1

By the end of this lecture, you will be able to

- describe the main components of any ML algorithm
- describe how linear regression works
- **describe how logistic regression works**

Logistic regression

```
In [13]: from sklearn.linear_model import LogisticRegression
LogReg = LogisticRegression() # initialize a simple logistic regression model
LogReg.fit(X_train,y_train) # we will learn what happens when you issue this line
```

- name is misleading, logistic regression is for classification problems!
- the model:

$$y'_i = \frac{1}{1+e^{-z}}, \text{ where}$$

$$z = w_0 + \sum_{j=1}^m w_j x_{ij}$$

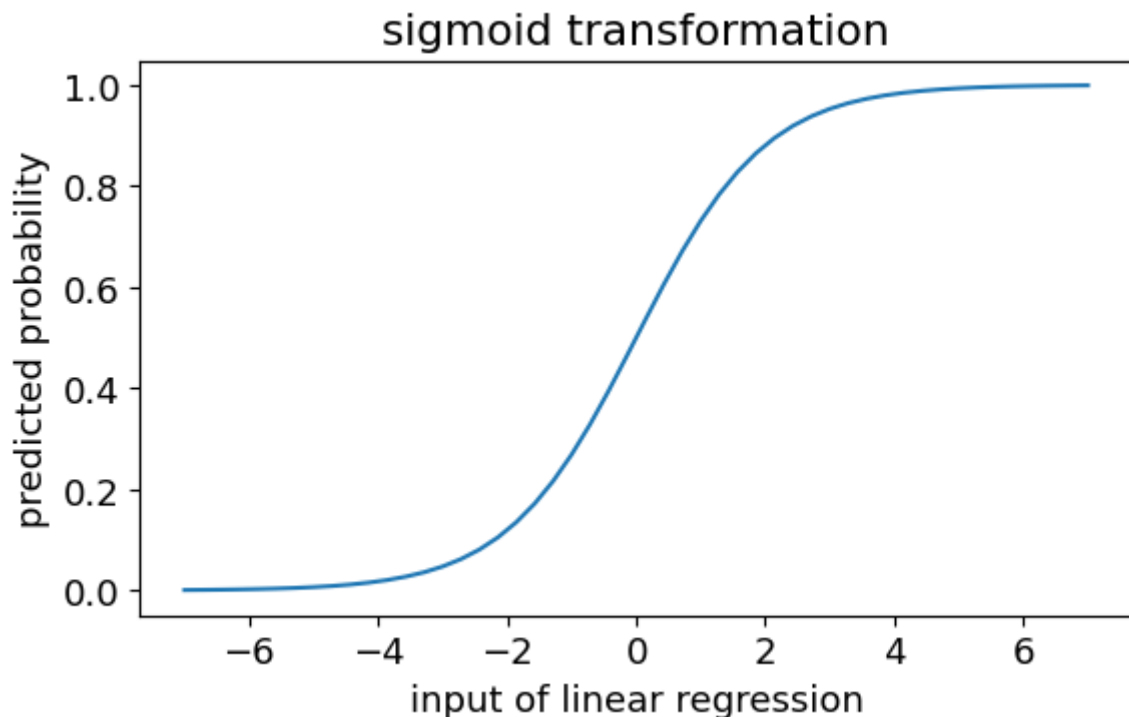
$f(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

- it is linear regression model but a sigmoid function is applied to its output

```
In [14]: def sigmoid(z):
          return 1/(1+np.exp(-z))

z = np.linspace(-7,7,50)

plt.figure(figsize=(6.4,3.6))
plt.plot(z,sigmoid(z))
plt.xlabel('input of linear regression')
plt.ylabel('predicted probability')
plt.title('sigmoid transformation')
plt.savefig('figures/sigmoid_trans.png',dpi=300)
plt.show()
```



The cost function

- the logloss metric is used as a cost function in logistic regression

$$L(w) = -\frac{1}{N} \sum_{i=1}^n [y_i \ln(y'_i) + (1 - y_i) \ln(1 - y'_i)]$$

$$L(w) = -\frac{1}{N} \sum_{i=1}^n \left[y_i \ln\left(\frac{1}{1+e^{-w_0+\sum_{j=1}^m w_j x_{ij}}}\right) + (1 - y_i) \ln\left(1 - \frac{1}{1+e^{-w_0+\sum_{j=1}^m w_j x_{ij}}}\right) \right]$$

Gradient descent

- the basic algorithm works but the `gradient_descent` function needs to be updated because the cost function changed!
- repeat until convergence:

$$w_j := w_j - l \frac{\partial L(w)}{\partial w_j},$$

where $\frac{\partial L(w)}{\partial w_j}$ is the gradient of the cost function at the current w location and l is the learning rate.

Mud card