Scientific computing and data visualization in Python

Adrian Kania[1]

[1]Department of Computational Biophysics and Bioinformatics

2025

Derivatives

derivative of f - function, describes the f (its monotonicity, enables to localize the maximum and minimum values)

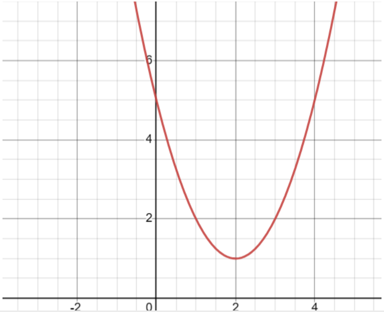
Example:

$$f(x) = x^2 - 4x + 5$$

then

$$f'(x) = 2x - 4$$

- $f'(x) < 0$ if $x < 2$ (f decreases)
- $f'(x) > 0$ if $x > 2$ (f increases)
- $f'(x) = 0$ if $x = 2$ (f has a minimum value here)



Derivatives

- $(x^n)' = nx^{n-1}$, for example: $(x^5)' = 5x^4$,
- $(ax)' = a$, for example: $(5x)' = 5$,
- $(a)' = 0$, for example: $(5)' = 0$,
- $(\sin x)' = \cos x$,
- $(\cos x)' = -\sin x$,
- $(e^x)' = e^x$.

Compund functions

- for $f(x) = x^4$, $f'(x) = 4x^3$,
- for $f(x) = (5x-4)^4$, $f'(x) = 4(5x-4)^3 \cdot 5$,
- for $f(x) = (x^3-5)^2$, $f'(x) = 2(x^3-5) \cdot 3x^2$.

Linear regression

Consider the following dependence:

- $3 \longrightarrow 7$,
- $2 \longrightarrow 5$,
- $-2 \longrightarrow -3$,
- $0 \longrightarrow 1$,

Questions:

- $1 \longrightarrow ?$,
- General rule $y(x) = ?$

Linear regression

Consider the following dependence:

- $3 \to 7$,
- $2 \to 5$,
- $-2 \to -3$,
- $0 \to 1$,

Questions:

- $1 \to 3$,
- General rule $y(x) = 2x + 1$

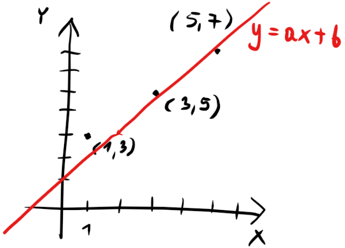
Unfortunately, practical examples are not such straightforward.

Linear regression

Let's say we have pairs of observations $(3, 5)$, $(4, 7)$ and $(1, 3)$. We want to find a linear dependence

$$y = ax + b$$

that suits the best to this data (not exact match but good approximation).



Linear regression

Let's say we have pairs of observations $(3, 5)$, $(4, 7)$ and $(1, 3)$. We want to find a linear dependence

$$y = ax + b$$

Consider the following expression (which refelcts the difference between observed y and predcited based on $ax + b$ relation)

$$L(a, b) = \frac{1}{3}[(5 - (3a + b))^2 + (7 - (4a + b))^2 + (3 - (1a + b))^2].$$

The aim is to minimize $L(a, b)$ - find a, b for which L have the minium value. We calculate derivatives for variables a and b.

- $\frac{\partial L}{\partial a} = \frac{1}{3}[2(5 - 3a - b) \cdot (-3) + 2(7 - 4a - b) \cdot (-4) + 2(3 - 1a - b) \cdot (-1)],$
- $\frac{\partial L}{\partial b} = \frac{1}{3}[2(5 - 3a - b) \cdot (-1) + 2(7 - 4a - b) \cdot (-1) + 2(3 - 1a - b) \cdot (-1)].$

Linear regression

Let's say we have pairs of observations $(3, 5)$, $(4, 7)$ and $(1, 3)$. We want to find a linear dependence

$$y = ax + b$$

We set derivatives to zero:

- $\frac{\partial L}{\partial a} = 2(5 - 3a - b) \cdot (-3) + 2(7 - 4a - b) \cdot (-4) + 2(3 - 1a - b) \cdot (-1) = 0,$
- $\frac{\partial L}{\partial b} = 2(5 - 3a - b) \cdot (-1) + 2(7 - 4a - b) \cdot (-1) + 2(3 - 1a - b) \cdot (-1) = 0.$

straight calculations result in linear system of equations:

- $13a + 4b = 23,$
- $8a + 3b = 15.$

and finally $a = \frac{9}{7}$ and $b = \frac{11}{7}$ (how do we know there is a minimum?). It means that

$$y = \frac{9}{7}x + \frac{11}{7} \approx 1.3x + 1.6$$

is the best linear approximation for these data.

Linear regression - 2D case

Consider the following observations $((1,2),4)$, $((5,6),-2)$, $((-3,2),1)$ and $((5,2),3)$. We want to find a linear dependence between y and (x_1, x_2).

$$y = w_1 x_1 + w_2 x_2 + w_0$$

where w_1, w_2 and w_0 are parameters we are searching for. Similarly to the previous case, we are considering the function

$$L(w_1, w_2, w_0) =$$
$$\frac{1}{4}[(4-(1w_1+2w_2+w_0))^2+(-2-(5w_1+6w_2+w_0))^2+(1-(-3w_1+2w_2+w_0))^2+(3-(5w_1+2w_2+w_0))^2].$$

By calculating $\frac{\partial L}{\partial w_1}$, $\frac{\partial L}{\partial w_2}$ and $\frac{\partial L}{\partial w_0}$, and set them to zero we may calculate w_1, w_2 and w_0.

Linear regression - 2D case

Consider the following observations $((1,2),4)$, $((5,6),-2)$, $((-3,2),1)$ and $((5,2),3)$. We want to find a linear dependence between y and (x_1,x_2).

$$y = w_1x_1 + w_2x_2 + w_0$$

where w_1, w_2 and w_0 are parameters we are searching for. Similarly to the previous case, we are considering the function

$$L(w_1,w_2,w_0) =$$
$$\frac{1}{4}[(4-(1w_1+2w_2+w_0))^2+(-2-(5w_1+6w_2+w_0))^2+(1-(-3w_1+2w_2+w_0))^2+(3-(5w_1+2w_2+w_0))^2].$$

By calculating $\frac{\partial L}{\partial w_1}$, $\frac{\partial L}{\partial w_2}$ and $\frac{\partial L}{\partial w_0}$, and set them to zero we may calculate w_1, w_2 and w_0.

Conclusion: If we have enough number of set of pairs $((x_1,x_2,...,x_n),y)$ we may find a linear depdendece $y = w_1x_1 + w_2x_2 + ... + w_nx_n + w_0$ that minimize the average difference between the real y and prediction (analytical solution exists). More precisely, $W = (X^TX)^{-1}X^TY$, where X - matrix of data $x_1,...,x_n$, Y - vector of y.

Proof - supplementary

$$L(W) = |XW - Y|^2$$

$$L(W) = (XW - Y)^T(XW - Y)$$

$$L(W) = Y^TY - Y^TXW - W^TX^TY + W^TX^TXW$$

then

$$\frac{\partial L(W)}{\partial W} = \frac{\partial(Y^TY - Y^TXW - W^TX^TY + W^TX^TXW)}{\partial W} = -2X^TY + 2X^TXW$$

setting the gradient to zero

$$-2X^TY + 2X^TXW = 0$$

results in

$$X^TXW = X^TY$$

and finally

$$W = (X^TX)^{-1}X^TY$$

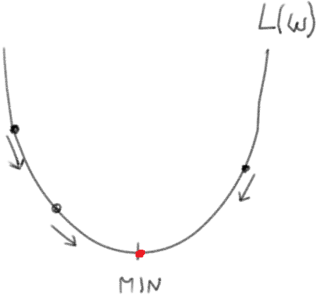
Some exercises

- Implement a linear model for 1D case ($y = ax + b$) according to the following dependencies:
$$a = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} \text{ and } b = \overline{y} - a\overline{x}$$
- Implement a linear model in general case according to:
$$W = (X^T X)^{-1} X^T Y$$
- Given $(3, 5)$ $(4, 6)$, $(5, 9)$ and $(6, 7)$ find the linear curve $y = ax + b$ that approximate the data. Visualize the data and model.
- Given $((3, 2), 1)$, $((4, -2), 7)$, $((5, 1), 3)$ and $((-2, 3), -4)$ find the linear dependence $y = w_1 x_1 + w_2 x_2 + w_0$ that approximate the data the best.

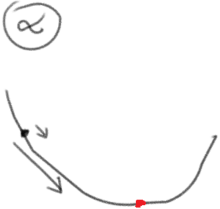
Gradient Descent

Sometimes, it is not easy to solve a system of equations (especially when they are not linear). There is an another method for finding parameters that **minimize** a given function.

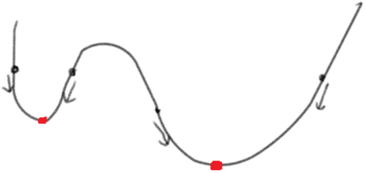
$$w_{new} = w_{old} - \alpha \frac{\partial L}{\partial w}.$$



SCHEME



ALPHA IMPACT



LOCAL MINIMUM

Gradient Descent for Linear Regression

Consider once again, the following observations $((1,2),4)$, $((5,6),-2)$, $((-3,2),1)$ and $((5,2),3)$. We want to find a linear dependence between y and (x_1, x_2).

$$y = w_1 x_1 + w_2 x_2 + w_0$$

where w_1, w_2 and w_0 are parameters we are searching for. The algorithm is the following:

- Start with some random parameters w_1, w_2 and w_0.
- Set α (for example $\alpha = 0.01$).
- Calculate $\frac{\partial L}{\partial w_1}$, $\frac{\partial L}{\partial w_2}$ and $\frac{\partial L}{\partial w_0}$.
- Update parameters w_1, w_2, w_0 according to:

$$w_{new} = w_{old} - \alpha \frac{\partial L}{\partial w}$$

- Repeat two last steps till converge.

Is linear model appropriate?

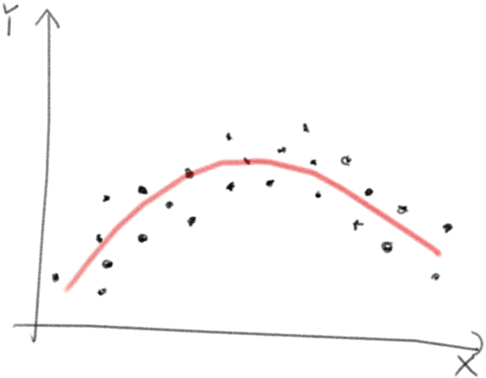
- **Pearson correlation coefficient** - measures the linear correlation
$$r = \frac{cov(x,y)}{s(x) \cdot s(y)}$$

If we have a model we may use

- **Loss function** - introduced earlier, measures the average square error between predicted and real value
$L(w_1, w_2, ..., w_0) = \frac{1}{n} \sum_i (y_i^{pred} - y_i)^2$ This function is alsko know ans **MSE** (mean square error).
- **Coefficient of determination** - measure of the goodness of fit of a model
$$R^2 = \frac{\sum_i (y_i^{pred} - \overline{y})^2}{\sum_i (y_i - \overline{y})^2}$$
or
$$R^2 = 1 - \frac{\sum_i (y_i^{pred} - y_i)^2}{\sum_i (y_i - \overline{y})^2}.$$

Non linear dependence

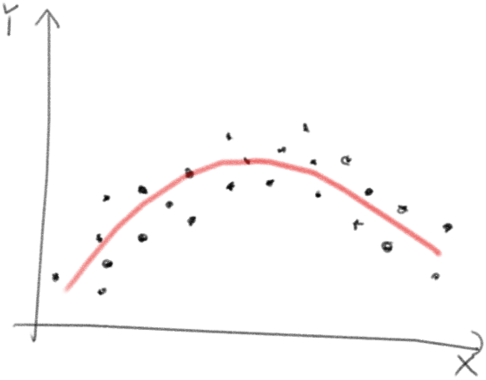
Sometimes, we observe more intricate dependence between variables x and y. In this case, simple linear approximation is illegitimate.



However, linear regression may be still useful. How?

Non linear dependence

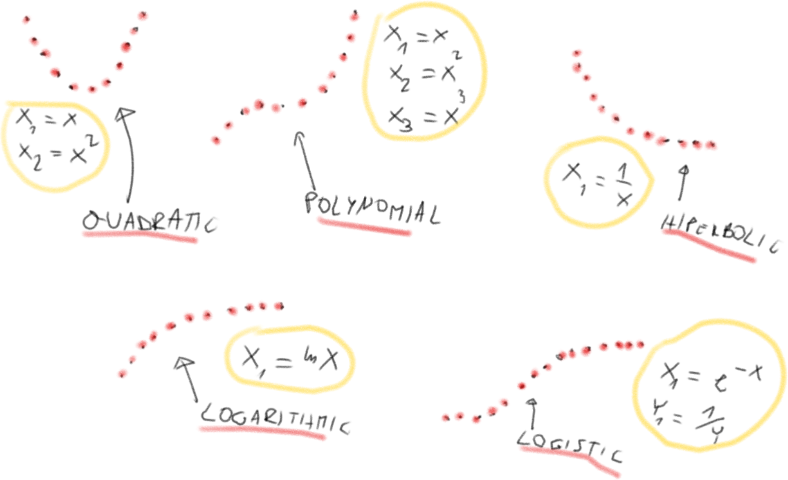
Sometimes, we observe more intricate dependence between variables x and y. In this case, simple linear approximation is illegitimate.



However, linear regression may be still useful. We may introduce new variables. Instead of using only y and x, we may consider $x_1 = x$ and $x_2 = x^2$ and then looking for:

$$y = w_1 x_1 + w_2 x_2 + w_0.$$

Transformations



In general, both Y and X may be transformed, and then linear model is fitted to $f(Y)$ and $g(X)$. Example: $Y^3 = aX^2 + bX + c$.

Exercise

- Using the quadratic transformation, fit the model to the data
    - $x = 1, 2, 3, 4, 5, 6$
    - $y = -1.68, -0.27, 3.93, 12.64, 25.91, 43.11$

  Visualize the data and model.

Regularization

To avoid large paramteres $w_1, w_2, ..., w_0$ which may results in overfitting, regularization is applied.

- **Ridge Regression (L2)**

$$L^2 = \beta \sum_i w_i^2$$

- **Lasso Regresssion (L1)**

$$L^1 = \beta \sum_i |w_i|$$

and finally new cost function is:

$$L = \frac{1}{n} \sum_i (y_i - y_i^{pred})^2 + L^1 \text{ or } L = \frac{1}{n} \sum_i (y_i - y_i^{pred})^2 + L^2$$

Logistic regression

Assume we have a variable Y which takes only two values - 0 and 1 (categories). Based on features $X_1, X_2, ..., X_M$ we want to predict the probability of occuring category 1. More precisely, we want to approximate $p = P(Y = 1)$. Notice that p is a function of variables $X_1, X_2, ..., X_M$. We are searching for a function which values are between 0 and 1 - may be interpreted as a probability. An example is

$$f(x) = \frac{1}{1+e^{-x}}$$

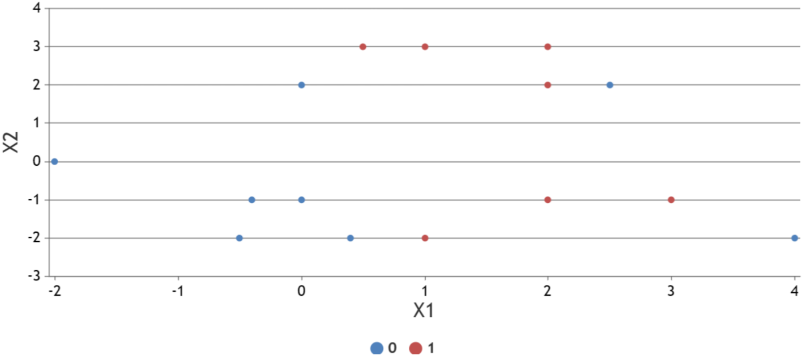
Finally, $P(Y = 1)$ may be modelled as:

$$\frac{1}{1+e^{-(w_1x_1+w_2x_2+...+w_0)}}$$

This expression will be denoted later as y_{pred}.

Example

X1	X2	Y
0	-1	0
0,4	-2	0
-0,5	-2	0
-0,4	-1	0
4	-2	0
-2	0	0
2,5	2	0
0	2	0
0,5	3	1
1	3	1
3	-1	1
1	-2	1
2	3	1
2	2	1
2	-1	1



Example

- Random (initial) parameters

$$w_1 = w_2 = w_0 = 0$$

For $(0, -1)$

  - $P(Y = 1) = \frac{1}{1+e^{-(0 \cdot 0 + 0 \cdot (-1) + 0)}} = \frac{1}{1+e^0} = \frac{1}{1+1} = \frac{1}{2}$,
  - $P(Y = 0) = 1 - \frac{1}{2} = \frac{1}{2}$

- After the learning step

$$w_1 = 0.5861, \; w_2 = 0.4331, \; w_0 = -0.9011$$

For $(0, -1)$

  - $P(Y = 1) = \frac{1}{1+e^{-(0.5861 \cdot 0 + 0.4331 \cdot (-1) - 0.9011)}} = \frac{1}{1+e^{1.3342}} = \frac{1}{1+3.8} = 0.21$,
  - $P(Y = 0) = 1 - 0.21 = 0.79$ (more probable)

How to compare two distributions

Cross Entropy

$$H(P,Q) = -\sum_x p(x) \log q(x)$$

Properties:

- $H(P,Q) \geqslant H(P,P)$,
- $H(P,Q) = H(P,P)$ only if $P = Q$.

Example:

X	-1	2	3
P	0.2	0.3	0.5
Q	0.6	0.2	0.2

$$H(P,Q) = -(0.2 \log 0.6 + 0.3 \log 0.2 + 0.5 \log 0.2) = 0.6$$

$$H(P,P) = -(0.2 \log 0.2 + 0.3 \log 0.3 + 0.5 \log 0.5) = 0.45$$

Binary Cross-Entropy (BCE)

If random a variable Y describes the category, it posseses only two values: 0 and 1. If true probability $y = P(Y = 1)$ and $1 - y = P(Y = 0)$ and model $y_{pred} = P(Y = 1)$ and $1 - y_{pred} = P(Y = 0)$. Then:

$$H(y, y_{pred}) = -(y \log y_{pred} + (1 - y)\log(1 - y_{pred}))$$

Now, we consider n data which has assigned categories $y_1, y_2, \ldots .y_n$. Our model predicts the probabilities y_1^{pred}, y_2^{pred}, ..., y_n^{pred}. To validate our model we should have

$$BCE = -\frac{1}{N}\sum_i y_i \log(y_i^{pred}) + (1 - y_i)\log(1 - y_i^{pred})$$

as small as possible. To find optimal parameters we use gradient descent method

$$w_{new} = w_{old} - \alpha\frac{\partial BCE}{\partial w}.$$

More than two categories

$$P(Y = k) = \frac{e^{z_k}}{\sum_j e^{z_j}}$$

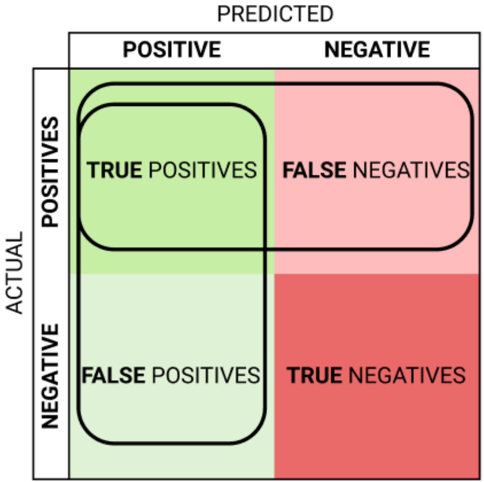
where $z_s = \sum_k w_{sk} x_k + w_{s0}$.

Loss function

$$CE = -\frac{1}{n} \sum_i \sum_k y_{ik} \log(y_{ik}^{pred})$$

where $k = 1, ..., M$ (number of categories).

Model evaluation



Model evaluation

- **Accuracy**

$$A = \frac{TP+TN}{TP+FN+TN+FP}$$

- **Sensitivity**

$$S_n = \frac{TP}{TP+FN}$$

- **Specificity**

$$S_p = \frac{TN}{TN+FP}$$

- **Precision**

$$P = \frac{TP}{TP+FP}$$

- **Recall**

$$R = \frac{TP}{TP+FN}$$

- **F-score**

$$F_1 = \frac{2}{P^{-1}+R^{-1}} = \frac{2TP}{2TP+FP+FN}$$

Model evaluation - example

Suppose we have the following actual data and predictions (0 - negative, 1 - positive).

Observation	Actual y	Predicted y
1	0	0
2	0	0
3	0	0
4	0	1
5	0	0
6	1	1
7	1	0
8	1	0
9	1	1
10	1	1

Calculate: A, S_n, S_p, P, R and F_1.

Model evaluation - example

Suppose we have the following actual data and predictions (0 - negative, 1 - positive).

Observation	Actual y	Predicted y
1	0	0
2	0	0
3	0	0
4	0	1
5	0	0
6	1	1
7	1	0
8	1	0
9	1	1
10	1	1

- $TP = 3$,
- $TN = 4$,
- $FP = 1$,
- $FN = 2$.

We may change the critical value

Till now, we assumed that if $p > 0.5$ we choose the first (1) category. But in general, we may change this critical value. It may be beneficial, especially, when we want to have one category predicted better.

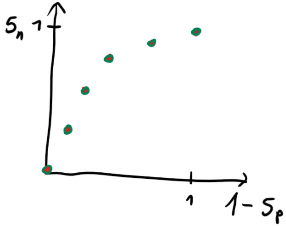
Observation	Actual y	y_pred	Predicted y if pc = 0.5	Predicted y if pc = 0.4
1	0	0.03	0	0
2	0	0.10	0	0
3	0	0.44	0	1
4	0	0.52	1	1
5	0	0.23	0	0
6	1	0.74	1	1
7	1	0.49	0	1
8	1	0.47	0	1
9	1	0.85	1	1
10	1	0.90	1	1

In the second case, all 1-category observations are found.

We may change the critical value

Till now, we assumed that if $p > 0.5$ we choose the first (1) category. But in general, we may change this critical value. It may be beneficial, especially, when we want to have one category predicted better.

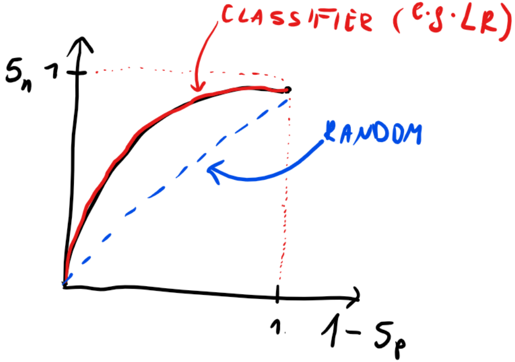
Observation	Actual y	y_pred	Predicted y if pc = 0.5	Predicted y if pc = 0.4
1	0	0.03	0	0
2	0	0.10	0	0
3	0	0.44	0	1
4	0	0.52	1	1
5	0	0.23	0	0
6	1	0.74	1	1
7	1	0.49	0	1
8	1	0.47	0	1
9	1	0.85	1	1
10	1	0.90	1	1

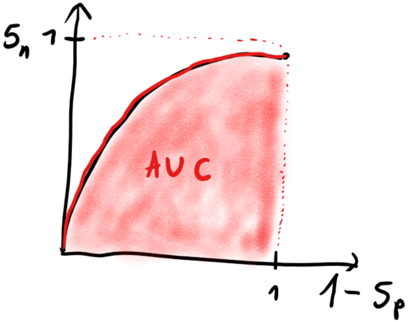


In the second case, all 1-category observations are found.

ROC curve

ROC - dependence between $1 - S_p$ and S_n.

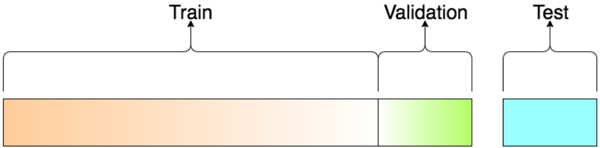




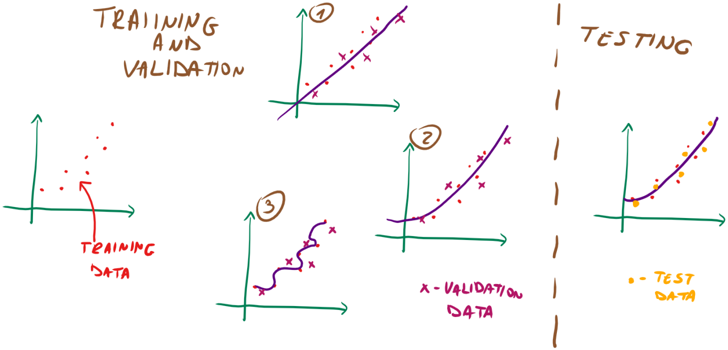
Questions

- Why all classifiers begin at $(0, 0)$ and end at $(1, 1)$.
- Where is the best classifier?

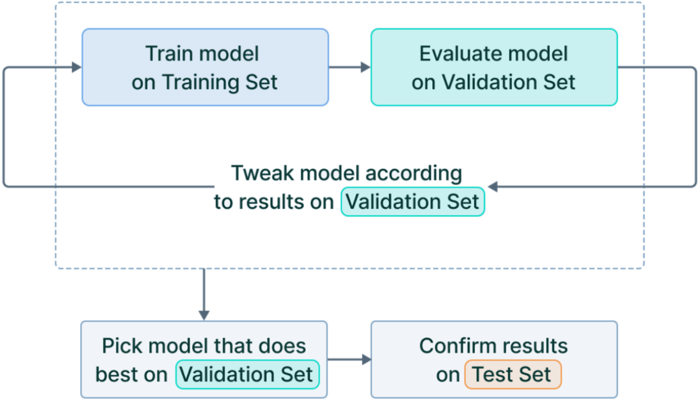
It's worth splitting the data



Validation data may be used to specify **hyperparameters**. In simple models, we divide data only into train and test sets.



It's worth splitting the data



Scientific paper reading

Consider the paper **Prediction of Gene Expression Patterns With Generalized Linear Regression Model**. Read at least the abstract and then, try to answer the following questions:

- What is the function of Oct4? (Introduction, page2)
- What did the authors model? Which variables? (Introduction, page2)
- What kind of data they used and where did they find them? (Materials and methods, page3)
- How Oct4 combination intensities were expressed? (Materials and methods, page3)
- How many cell development stages (days) were considered for gene analysis? (Materials and methods, page3)
- Which variables (Oct4 combination intensities) have the strongest correlation with expression levels of considered genes? (Table2, page4)
- How many models were considered? (page4).
- Which model was used to describe expression patterns of the Cnbp gene? What was the R^2 in this case? (page6; Table5, page7). According to the estimated parameters, which part of the model seems to have the lowest impact on the expression? Compare it with Table2. Make a comment on it.