Underfitting - Overfitting

Bias- Variance Tradeoff -2

↳ Linear Regression

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Underfitting-Overfitting.ipynb | Google

google.com/webhp?hl=en&ictx=2&sa=X&ved=0ahUKEwiA5JmH9qf3AhW5SmwGHYniBL0QPQgJ

Update

Gmail Images

Google Search I'm Feeling Lucky

It's Earth Day. See how Google is working to combat climate change

Google offered in: हिन्दी বাংলা తెలుగు मराठी தமிழ் ગુજરાતી ಕನ್ನಡ മലയാളം ਪੰਜਾਬੀ

India

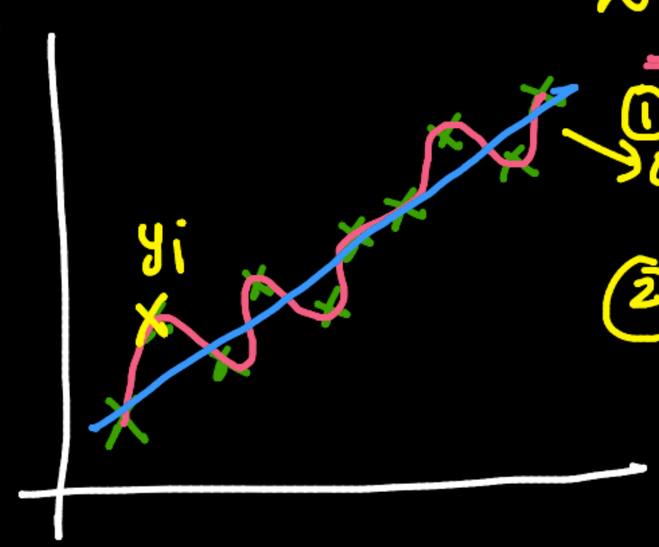
2 / 61

About Advertising Business How Search works Privacy Terms Settings

Agenda:

- Underfitting vs overfitting Tradeoff (examples)
- Intuition & Math (optimization)
- end-end Linear Regression
- Hyper-param toning
- code (simulation)

Recap



Train
① Occam's ~ : higher degree features
② generalizable : lower degree features f^1

red: Train Loss ≈ 0
blue: Train loss

$\text{Loss}^{\text{train}}_{\text{blue}} > \text{loss}^{\text{train}}_{\text{red}}$

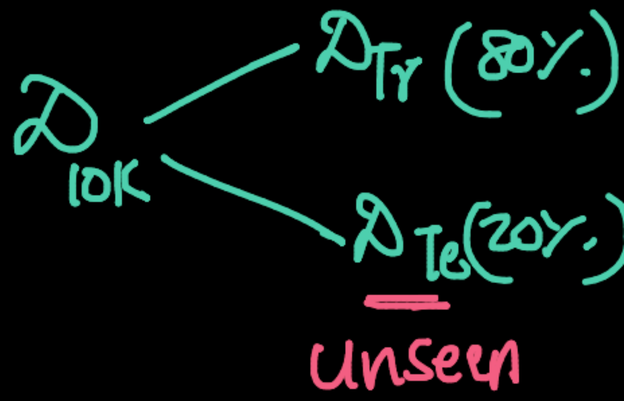
$f_1^1, f_1^2, f_1^3, \ldots$

$$\text{Loss} = \sum_{i=1}^{n} L_i \longrightarrow (y_i - \hat{y}_i)^2 \longrightarrow \text{squared loss (linear-reg)}$$

$$\min_{w_j} \sum_{i=1}^{n} \left[y_i - (w^T x_i + w_0) \right]^2$$

1. **Generalization**: ML

Train loss



$\mathcal{D}_{Tr}$ → Model

unseen { $\mathcal{D}_{Te}$ → Model → R-squared

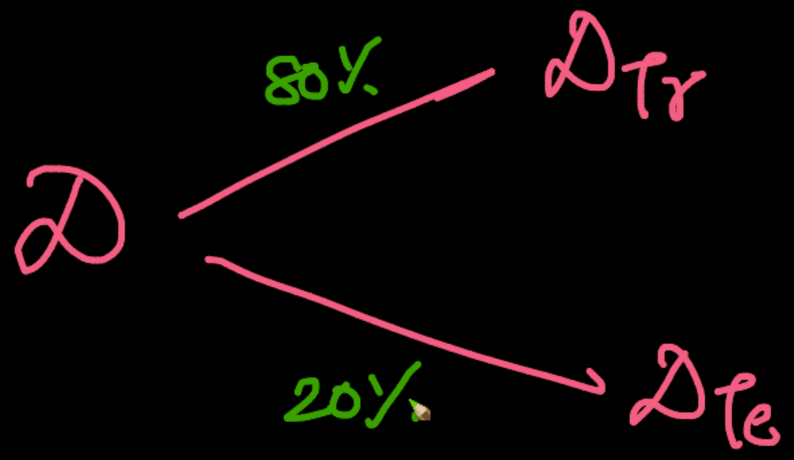
2. **Occam's razor**: science (ML)

↳ Pick the simplest model that fits/explains your data well!

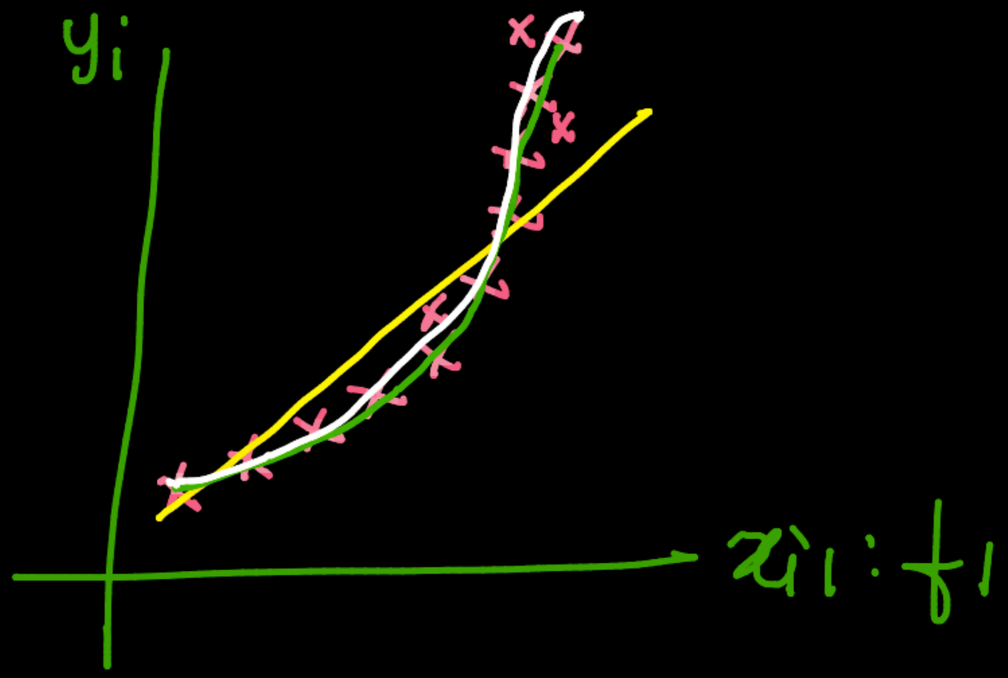
Chrome File Edit View History Bookmarks Profiles Tab Window Help

Real-world: All data is messy

7 / 7


80%
D_{Tr}
D
20%
D_{te}

Q



1. Linear : f^1
2. Quadratic : f_1, f_1^2
3. Cubic : f_1, f_1^2, f_1^3

simplest

✓ occam's razor

✓ [Gen: Test-error

Linear

#

Quad

L

complex

cubic

VL

4

bicubic

etc

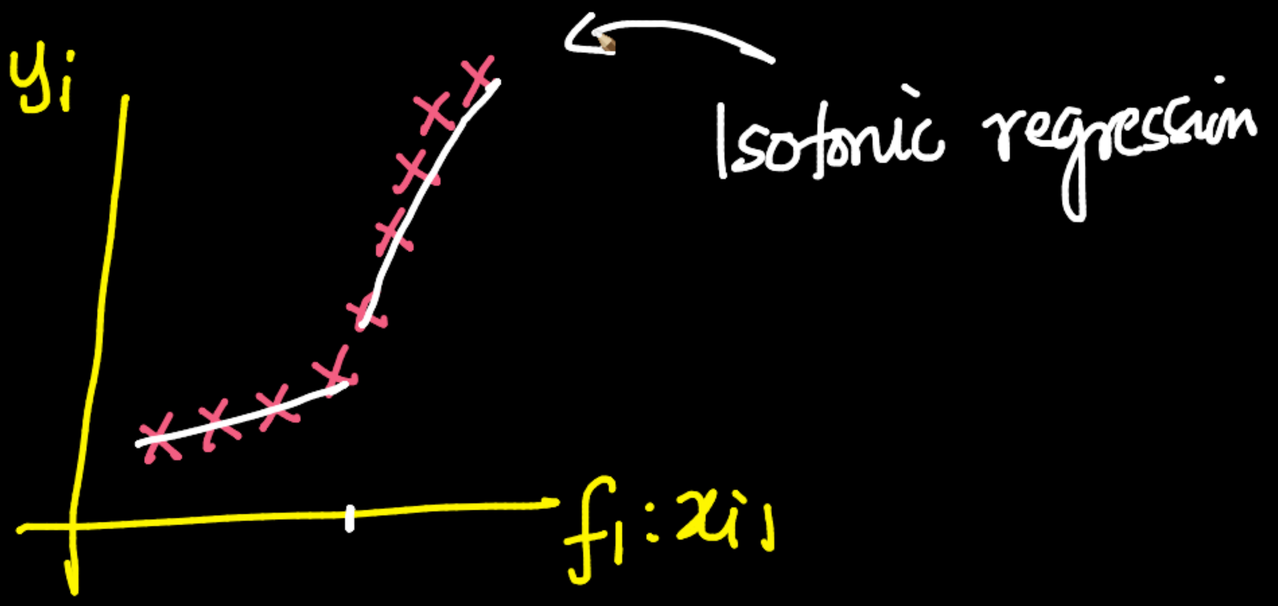
Linear-model: solving optmzn

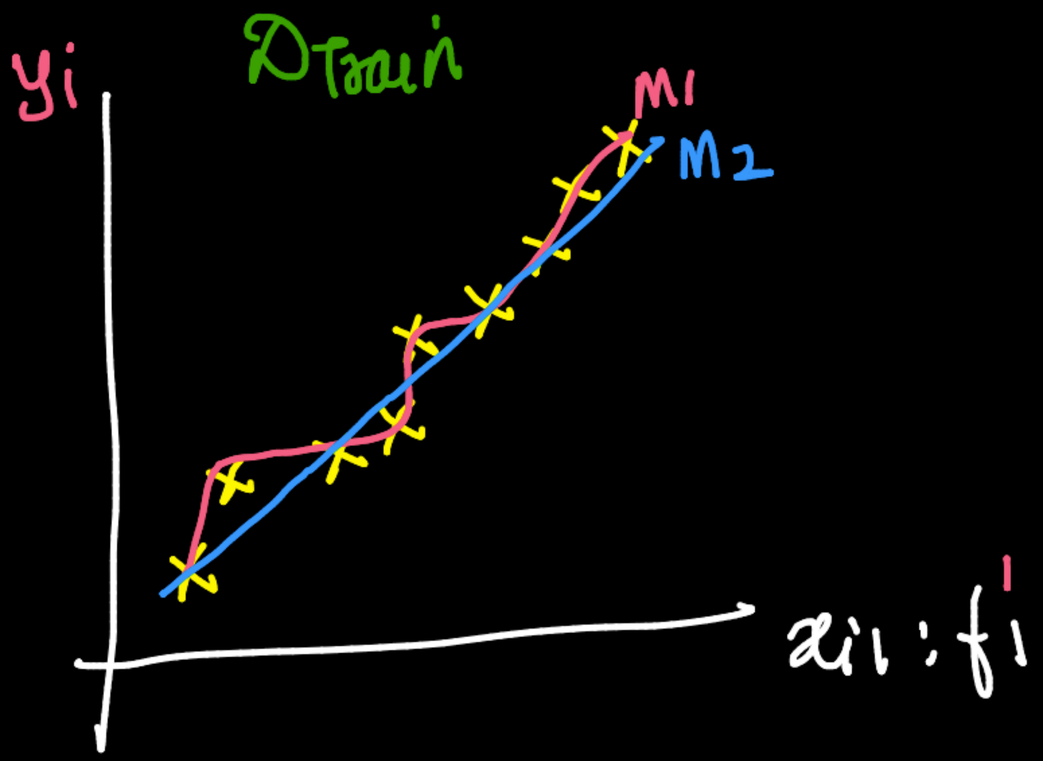
$$\min_{w_j} \sum_{i=1}^{n} \left[y_i - \left(\vec{w}^T x_i + \vec{w}_0 \right) \right]^2$$

(one-plane)

$f_1', f_2', \ldots$

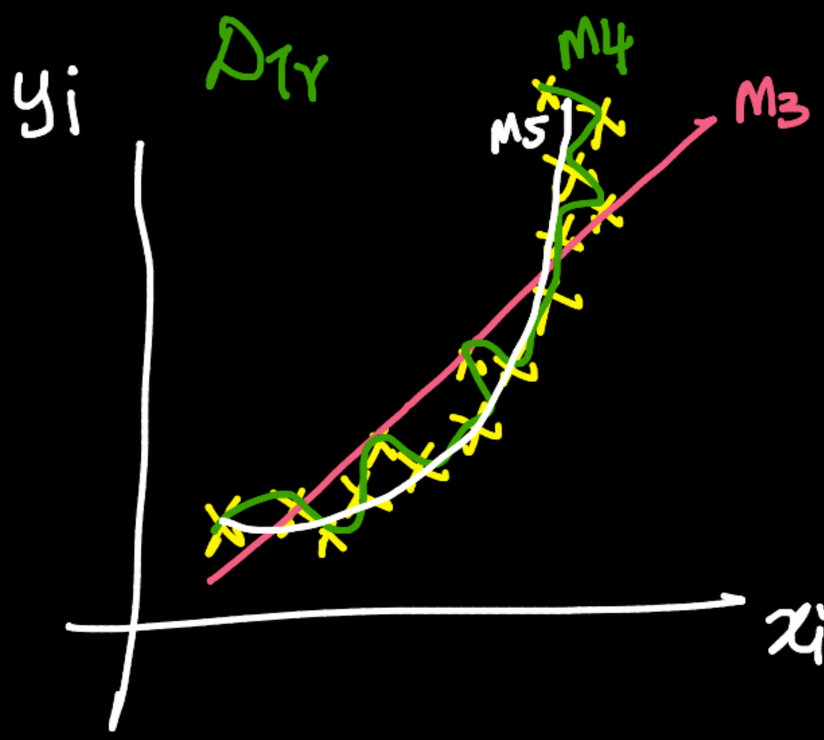
(no higher order variables)


yi
Isotonic regression
fi : xi



M1: overfit is D_{Tr}

Geom:



Linear
M_3
underfitting
not overfitting

higher order ← M_4 : overfitting

M_5 : not overfitting perfectly
Quad

Underfit
1st
Linear

perfectly
2nd
quad

overfit
higher

Simplest

[Underfit → DTr: 100 H, DTe: 120 H

{ perfectly → 20 Low, 30 Low
tradeoff

Complex

[overfit → 0.1 V.low, 50 not small

Plots → practice

D_{Tr}

D_{Te}

✓ (GOLD)

V·L

0·1

V·L

0·1

generalizati

Malik

{ overfit - Underfit Tradeoff

$$y_i = w_0 + w_1 x_{i1}^1 + w_2 x_{i1}^2 + w_3 x_{i1}^3 + w_4 x_{i1}^4$$

x_{i1} (original) f1

transformed features

Case 1:

M1 — overfit — $w_1, w_2, w_3, w_4 \neq 0$

M2 — $w_3 = w_4 = 0$ — $w_1, w_2 \neq 0$

M3 — underfit — $w_2 = w_3 = w_4 = 0$ — $w_1 \neq 0$

Math : weights on higher order poly $\neq 0$

↳ More chance of overfitting

R^2

↑

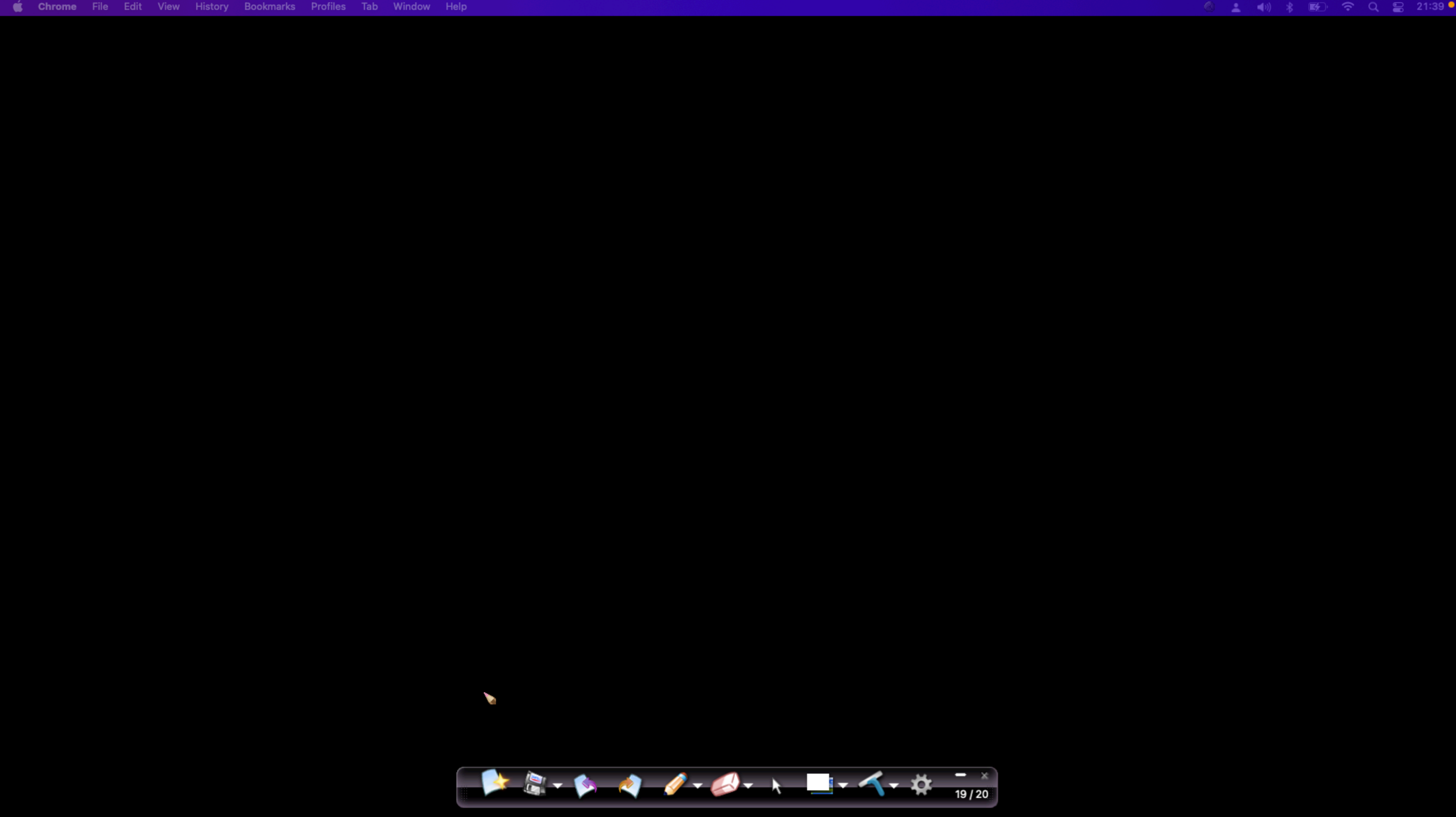
performance metric

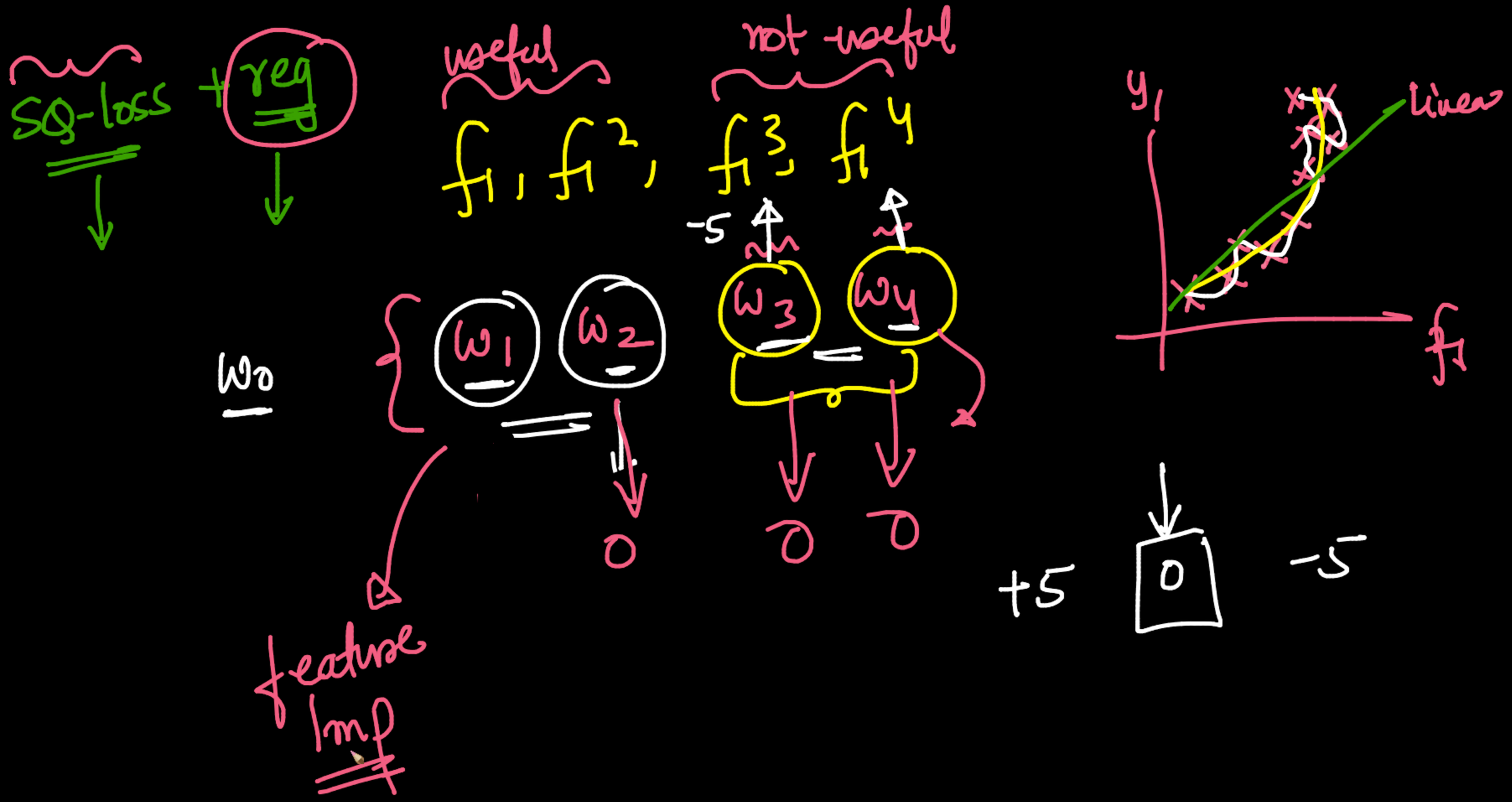
squared-loss

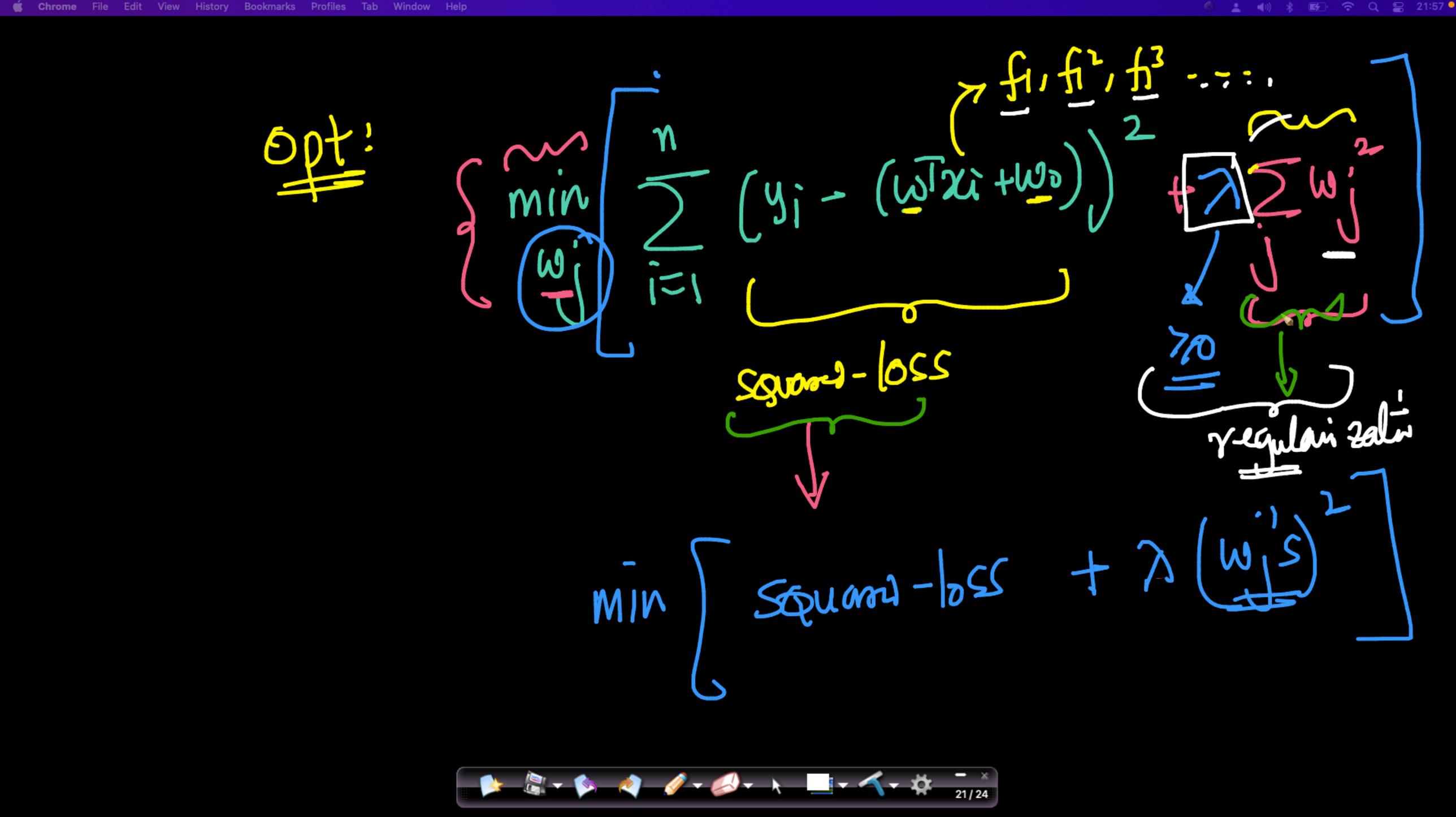
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

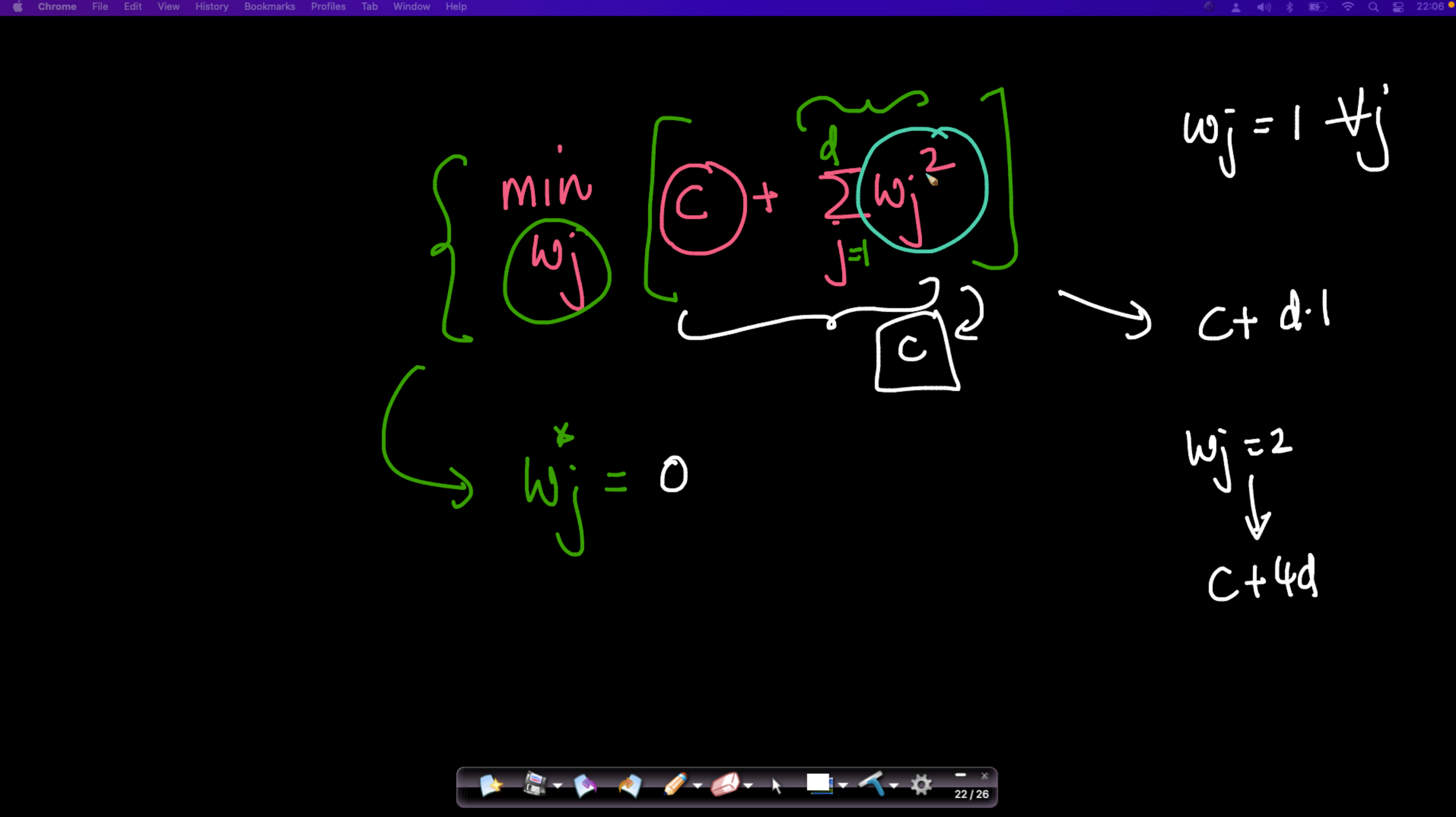
Loss → Minimize

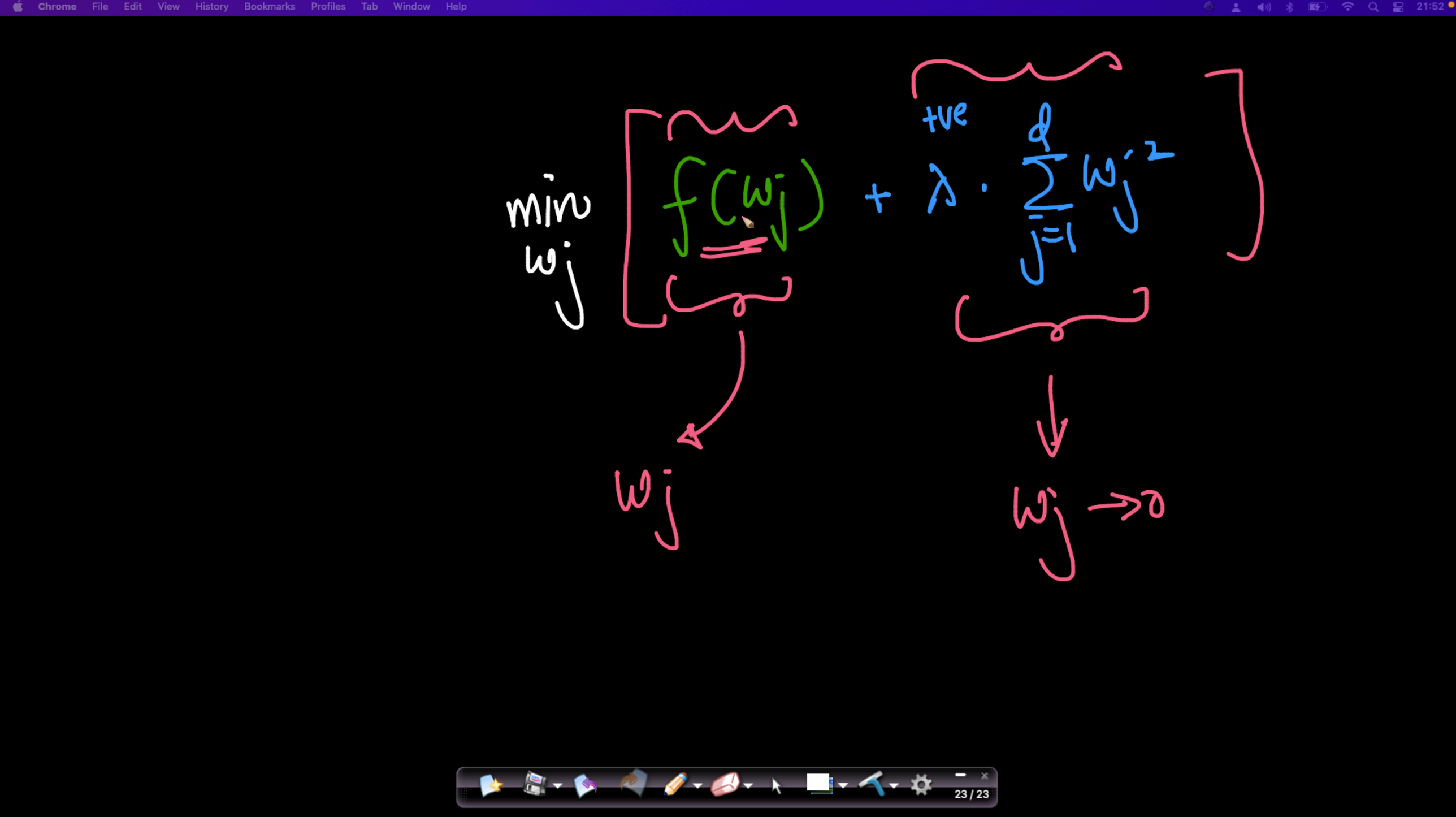
perf-metric → Maximize


Chrome File Edit View History Bookmarks Profiles Tab Window Help
21:39
19 / 20


SQ-loss
+ reg
useful
not useful
$f_1, f_1^2,$
f_1^3, f_1^4
-5
w_3
w_4
w_0
w_1
w_2
0
0
0
feature Imp
y_1
linear
f_1
+5
0
-5


Opt:
min
w_j
$\sum_{i=1}^{n}$
$(y_i - (w^T x_i + w_0))^2$
f_1, f_1^2, f_1^3 .....
$+\lambda \sum w_j^2$
$\lambda \geq 0$
regularization
squared-loss
$\min [\text{squared-loss} + \lambda (w_j's)^2]$


min
w_j
$c + \sum_{j=1}^{d} w_j^2$
c
$w_j = 1 \; \forall j$
$c + d \cdot 1$
$w_j^* = 0$
$w_j = 2$
$c + 4d$


min
wj
f(wj)
+
λ ·
+ve
wj → 0

$$\min_i \quad \overbrace{\text{squared loss}_i}^{} \quad + \quad \lambda \overbrace{\sum_j w_j^2}^{}$$

GD → $\frac{\partial \mathcal{L}}{\partial w_j}$

$\mathcal{L}$

$$\min_{w_j} \sum_i \text{SQ.loss}_i + \lambda \sum_j w_j^2$$

w_j^4

Q

$\lambda \sum_j |w_j|$ (abs) → L1 reg

① not diff @ $w_j = 0$

(hack (later))

L_2-reg

$w^T w$

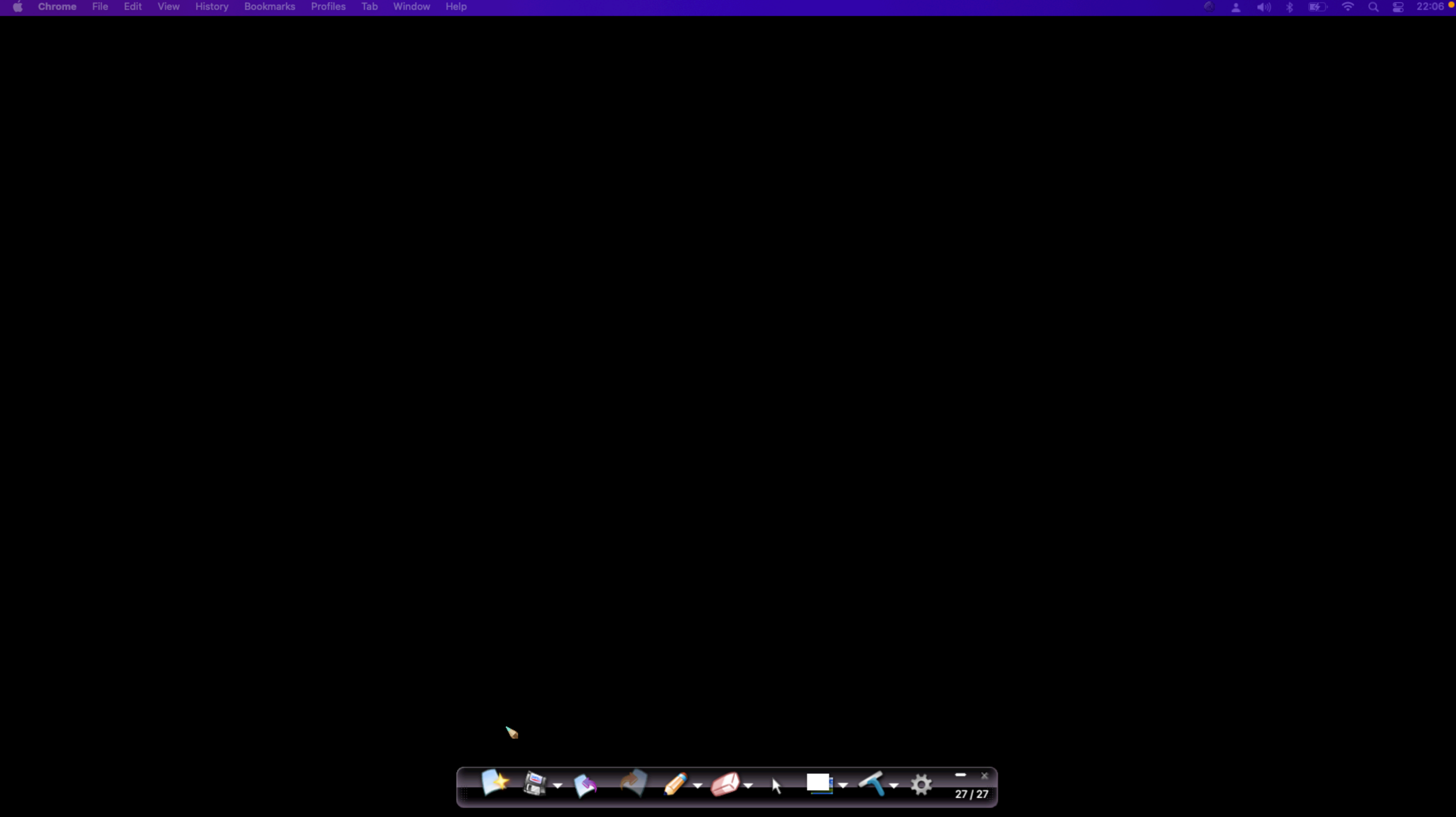
$= \|w\|_2 = L_2$ norm

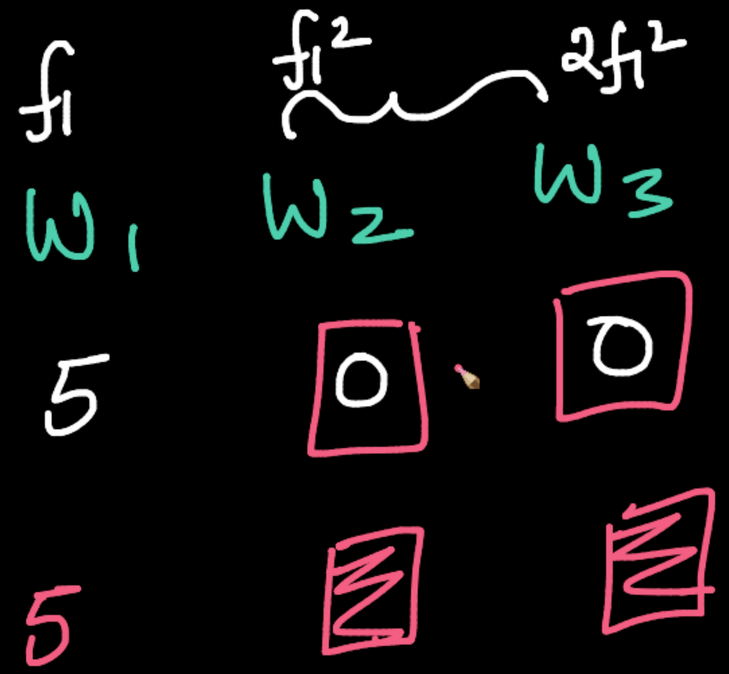
$w_j \rightarrow$ +ve, -ve

$w_j^3 \rightarrow$ +ve, -ve

w_j^4

post class → revise (re think & revise)


Chrome File Edit View History Bookmarks Profiles Tab Window Help
22:06
27 / 27



λ

$y_i = w^T x_i + w_0$

params

$$\min_{w_j} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum w_j^2$$

SQ. loss

$\hat{y}_i \downarrow 0$

$n \hookrightarrow 10000$

LB $\downarrow$

λ → simpler

hyperparam

reg

Case 1:

Case 2:

Case 3:

Case 4:

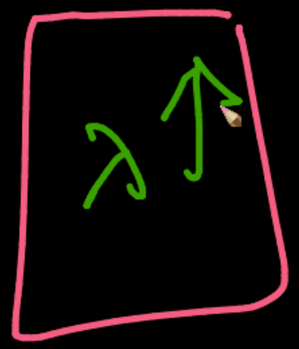
$\lambda = 0 \Rightarrow$ no reg → overfit

$\lambda = 10$

$\lambda = 100$

$\lambda = 1\text{Billion} \Rightarrow$ underfit

tradeoff



: overfit → underfit

↑

good-model

$$\min_{w_j} \sum_{i=1}^{n} \left(y_i - (w^T x_i + w_0) \right)^2 + \lambda \sum_j w_j^2$$

$\min_{w_j}$ → GD on each w_j

$\lambda \sum_j w_j^2$ → $\lambda w^T w$ → Constraint opt

objective

$w^T w = 1$

$\lambda (w^T w - 1)$

PCA

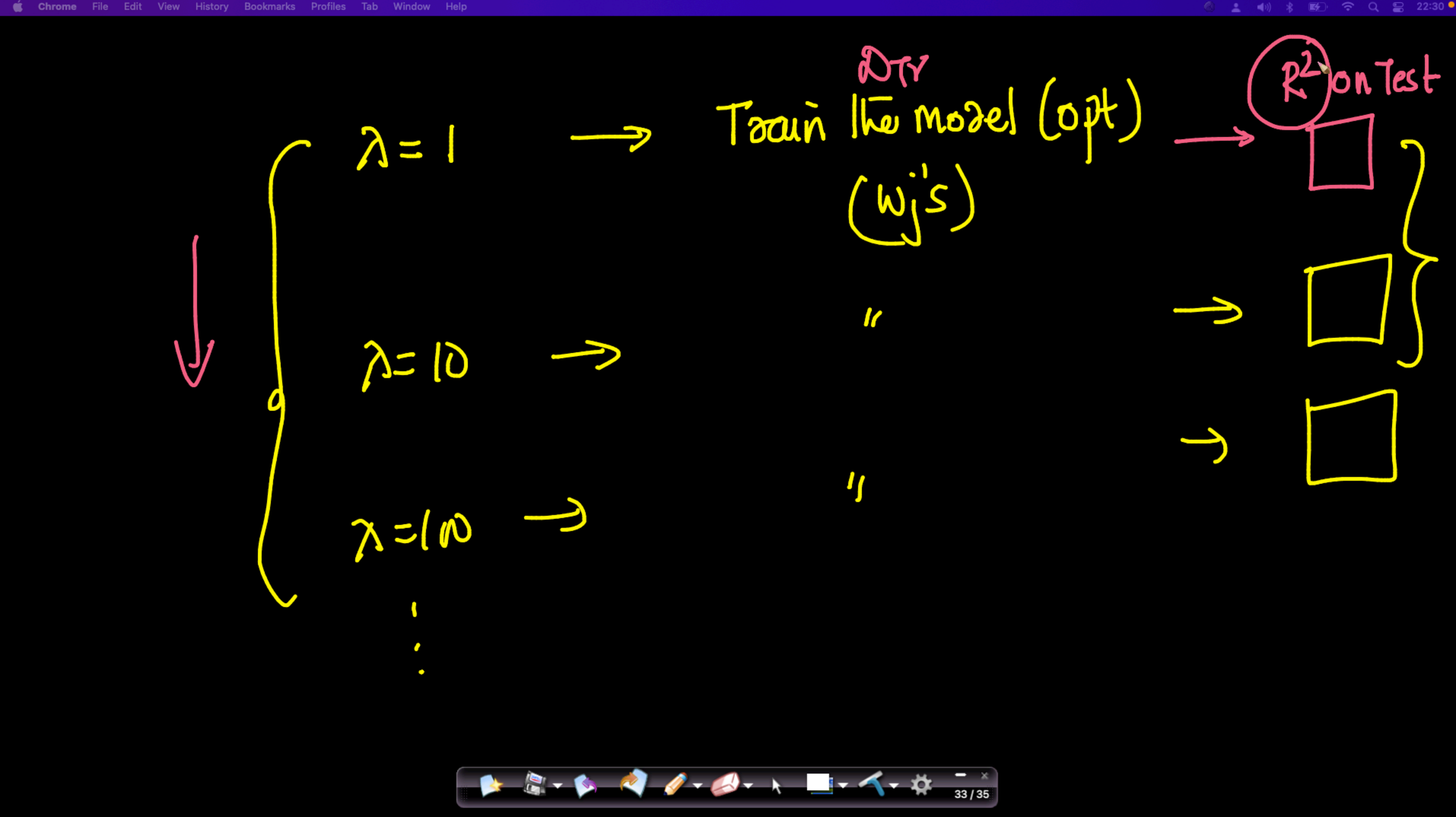
Constrained optimzn : $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$ → Yes

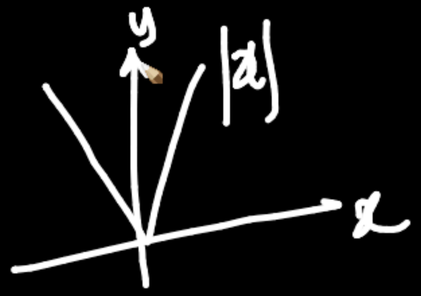
GD on λ to find x^*

ML-optimzn : NO

λ Control (tradeoff

(manual) → overfitting — underfit


Chrome File Edit View History Bookmarks Profiles Tab Window Help
22:30
DTr
Train the model (opt)
(wj's)
R2 on Test
λ = 1
λ = 10
λ = 100
"
"
33 / 35



$$\min_{w_j} \mathcal{L} \left\{ \sum_{i=1}^{n} \left[y_i - (w^T x_i + w_0) \right]^2 + \lambda \sum_j w_j^2 \right\}$$

$$\left\{ j \neq 0 : \quad \frac{\partial \mathcal{L}}{\partial w_j} = \sum_{i=1}^{n} 2 \left(y_i - (w^T x_i + w_0) \right) (-x_{ij}) + \lambda \cdot 2 \cdot w_j \right\}$$

$\lambda = 0$

$\lambda = 1$

100 feat

30 feat → 0

70 feat ≠ 0

overfit

Error

Test-error

Train-error

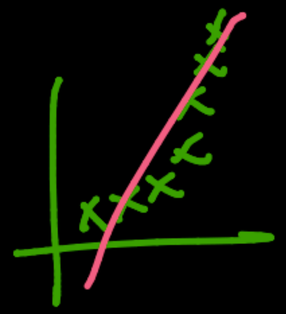
underfit

gen

0

λ

λ_{best}



Val-data

$$\min_{w_j} \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_j w_j^2$$

100 features → 70 feat ≠0 ; 30 =0 → some slightly useful features

min Loss + λ_B min weights

$\lambda_{Best} = 10$

↓

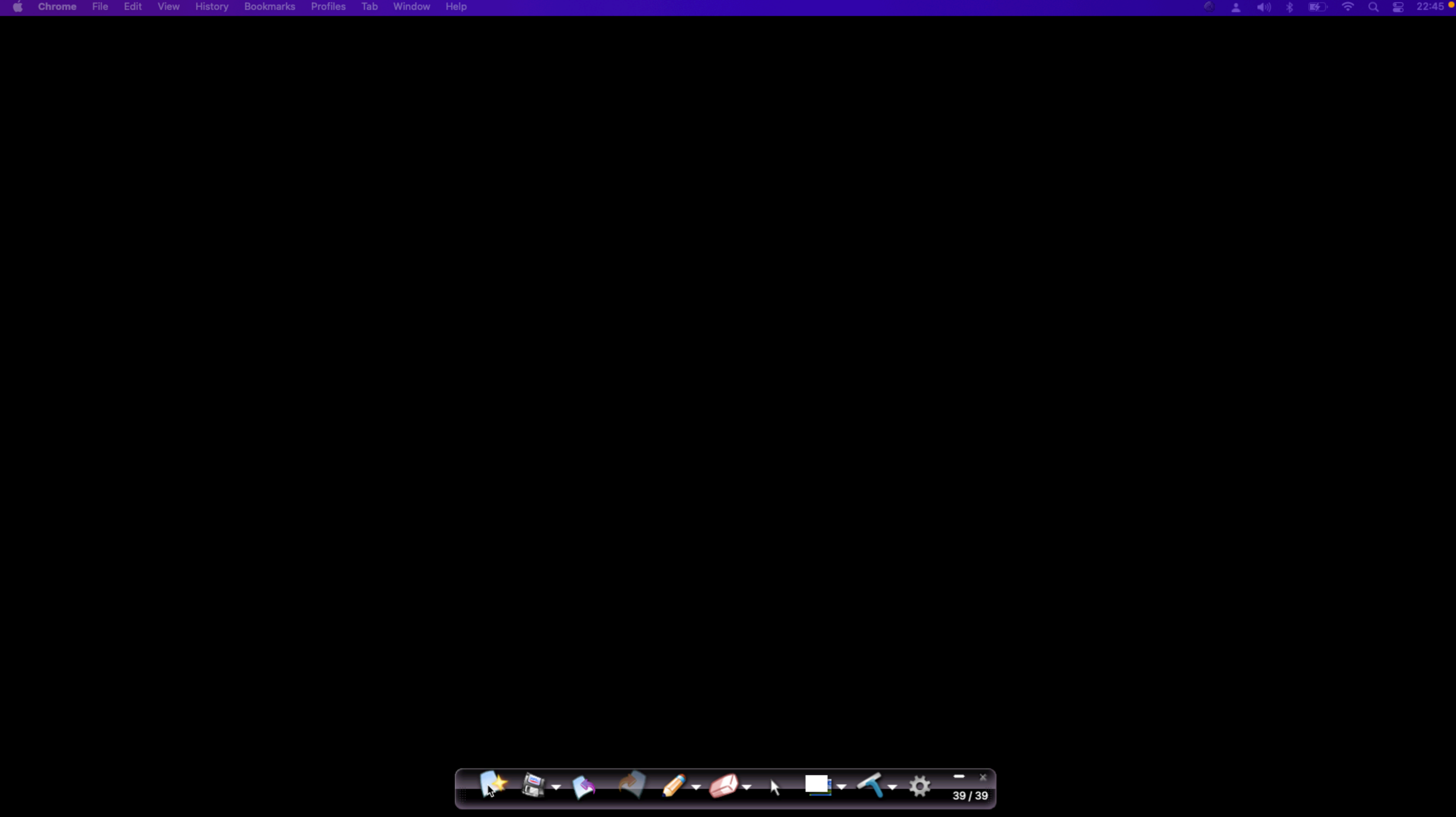
Test-data

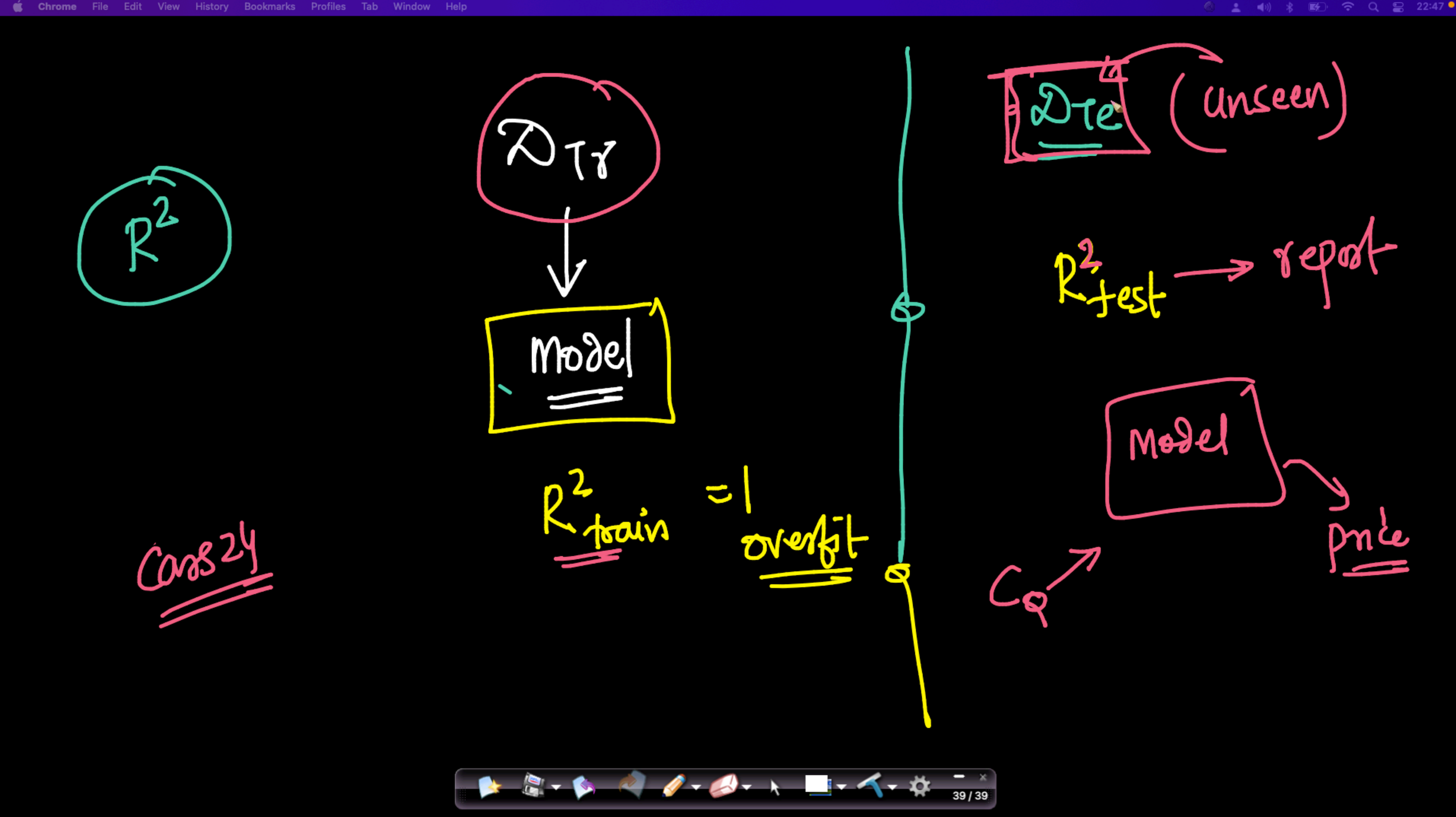
Unseen

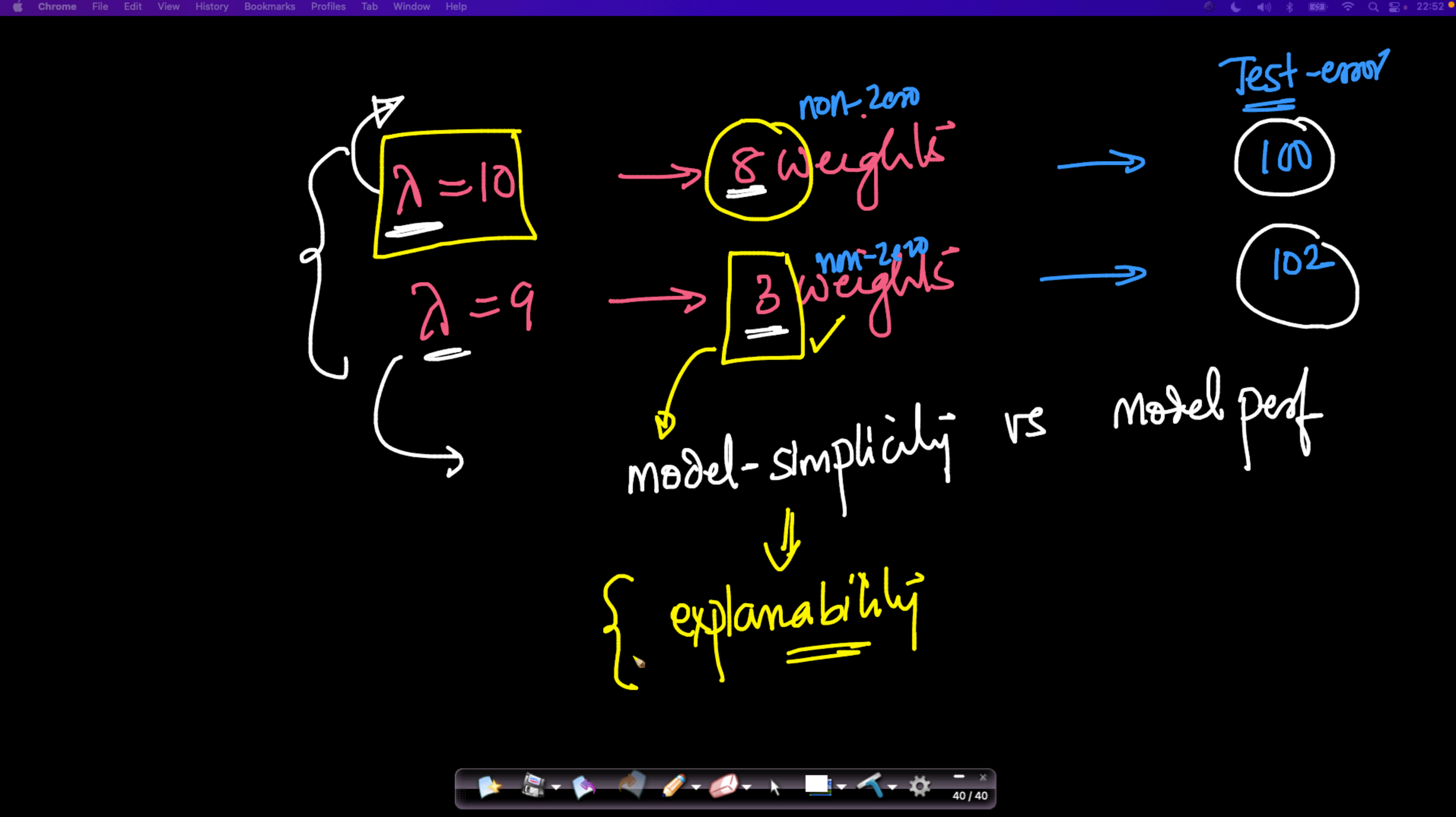
$\lambda = 1 \rightarrow$ Train the model $\rightarrow$ Test-error

$\Rightarrow \hat{w}_1 =$

$\lambda = 2 \rightarrow$ " $\Rightarrow \hat{w}_2 =$

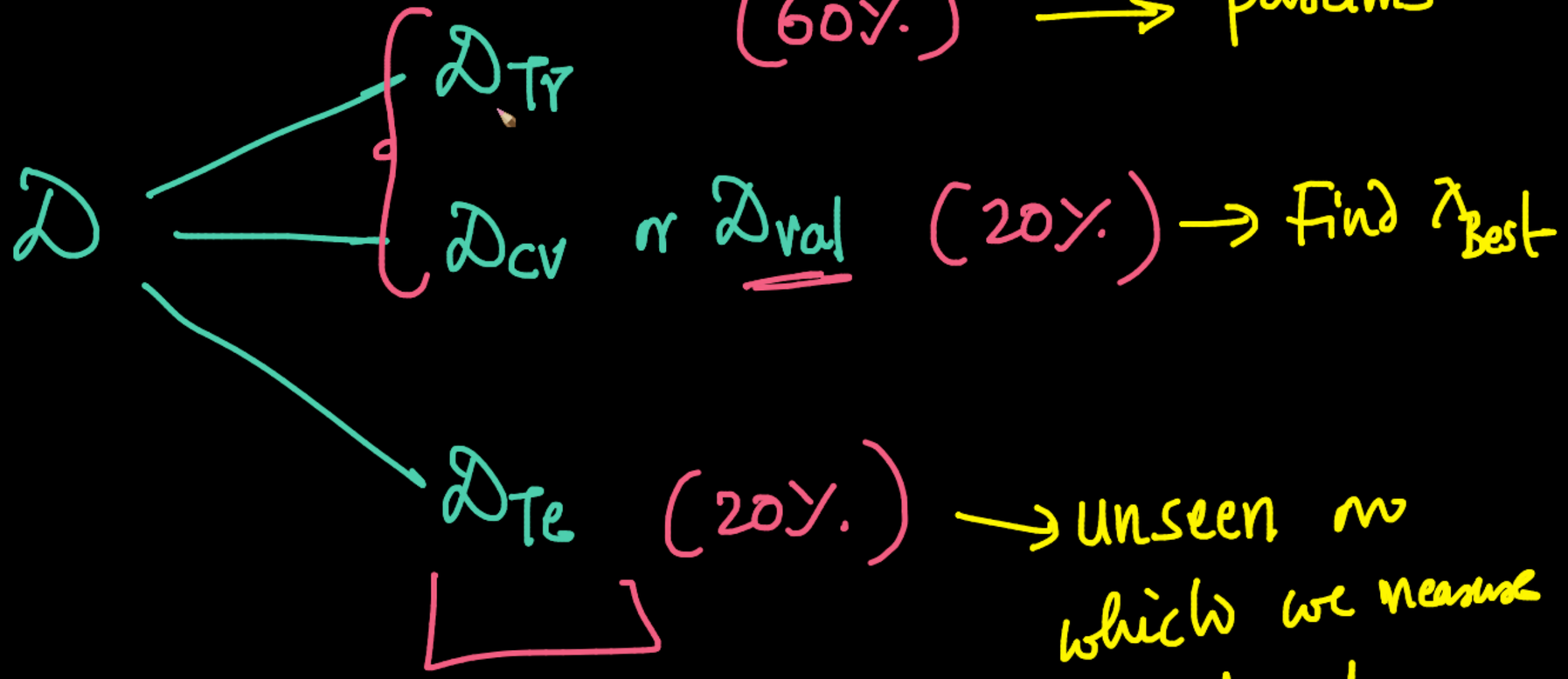

Chrome File Edit View History Bookmarks Profiles Tab Window Help
22:45
39 / 39


Chrome File Edit View History Bookmarks Profiles Tab Window Help
22:47
R^2
D_{Tr}
Model
R^2_{train} = 1
overfit
Cars24
D_{Te} (unseen)
R^2_{test} → report
Model
Price
39 / 39


Test-error
λ = 10
non-zero
8 weights
100
λ = 9
non-zero
3 weights
102
model-simplicity vs model perf
explainability

✓ Using test-data to find λ_{BEST} is not correct

Soln:


D
D_{Tr}
D_{CV} or D_{val}
D_{Te}
(60%.)
Find params
(20%.) → Find $\hat{}_{Best}$
(20%.)
→ unseen no which we measure model-perf

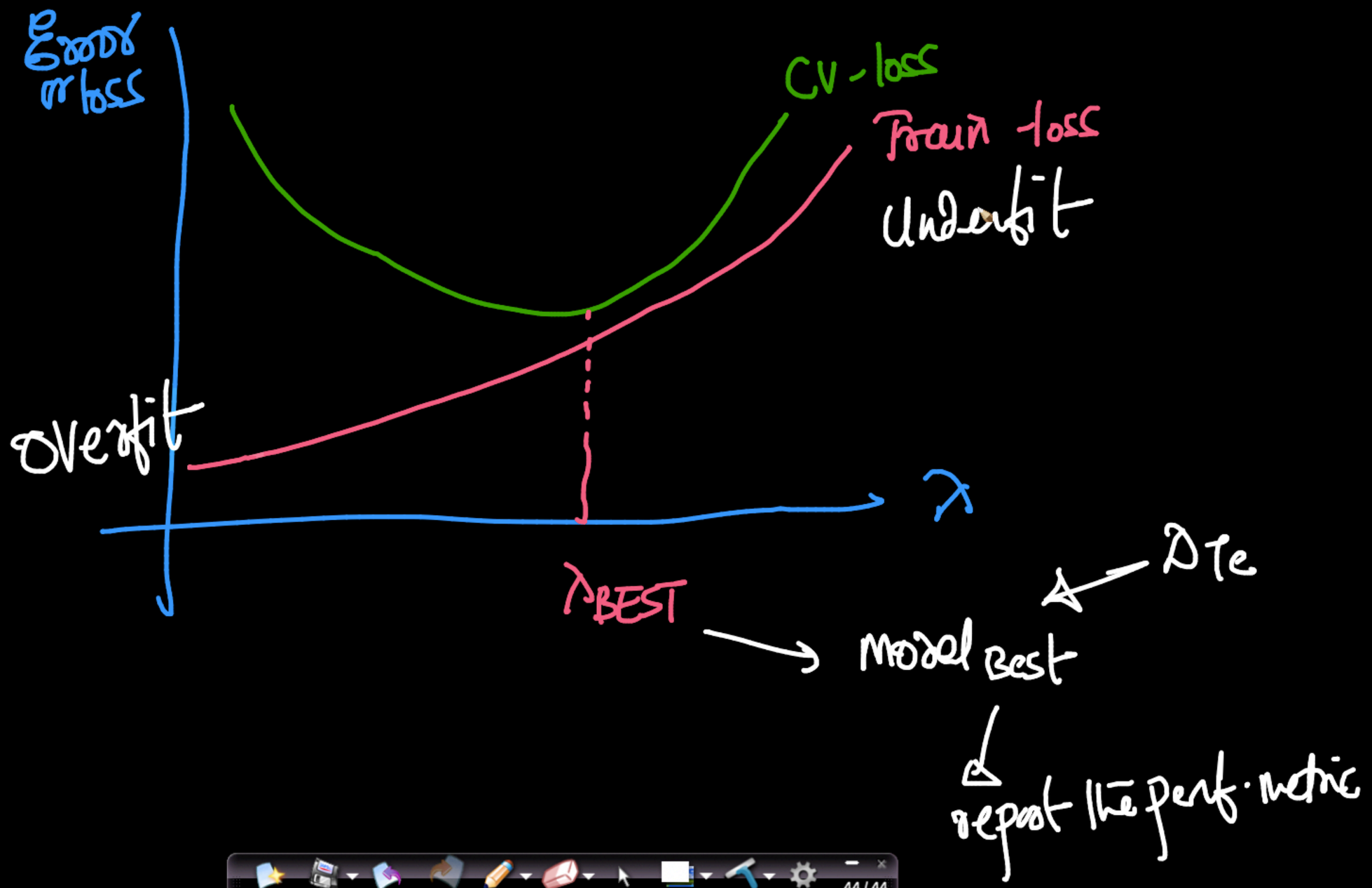
did not use D_{Te}

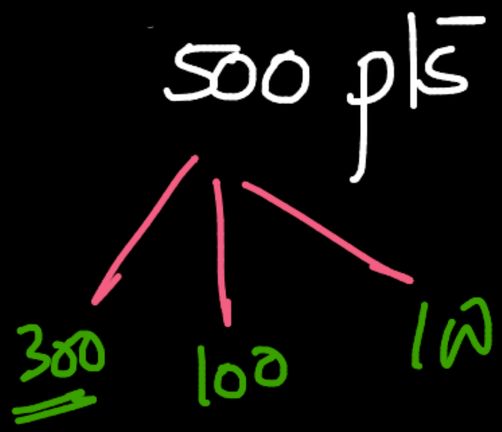
For $\lambda = 1, 10, 20, 50, 100, \ldots$

- Train a model (w_j's) on D_{Tr} with λ fixed
- Measure model-perf (R^2) on D_{CV}

- Pick the $\tilde{\lambda}_{Best}$ s.t it has lowest CV-error → $model_{Best}$

- Measure the perf of $model_{Best}$ on D_{Te} & report it


Error or loss
CV-loss
Train loss
underfit
overfit
λ
λBEST
Dte
model Best
report the perf. metric



features → 7

$d << n$

K-fold CV

Overfitting - Underfitting tradeoff

=

high Variance - high Bias tradeoff

later

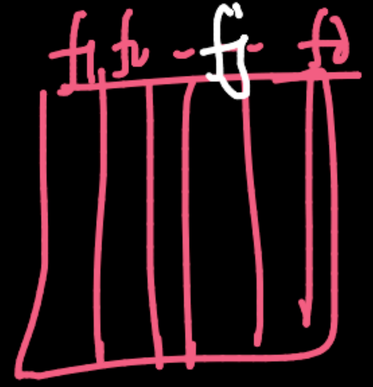
$y_i - \hat{y}_i = e_i$ → Bias + Variance (later) (stats)

→ disb

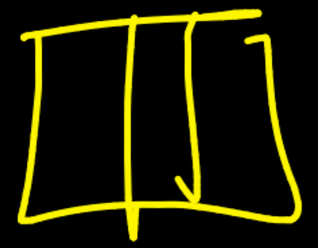
Expectations

$E(X)$

Summarize



1. D_{Tr}, D_{cv}, D_{te} (Best, unseen)
2. col. standardize D_{Tr} → $\forall j: \overline{m_j; s_j}$ (Store)
3. Using m_j & s_j → Transform D_{cv} & D_{te} Dk



4 For $\lambda = 1, 10, 20, \cdots$

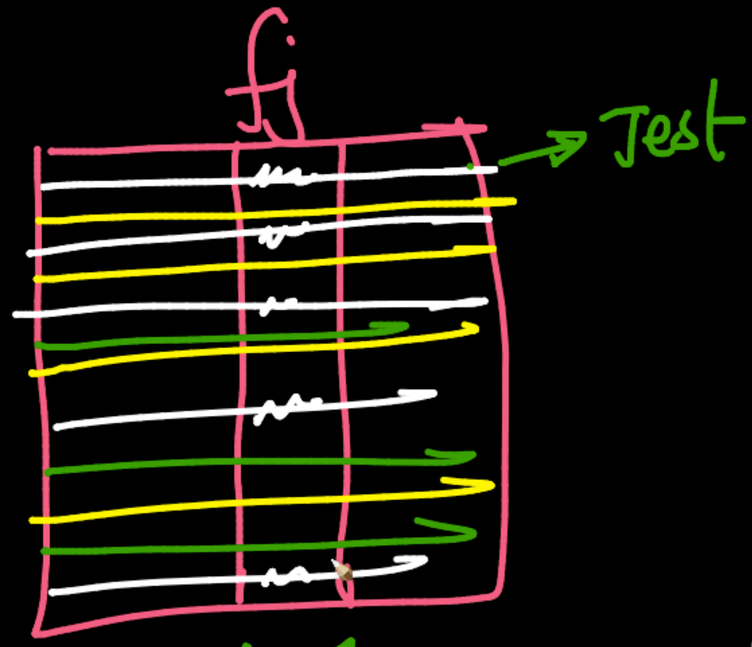
- Train model on Dtr

$$\sum_{i=1}^{train} \text{SQ.loss}_i + \lambda \sum_j w_j^2$$

→ GD

5 pick λ_{BEST} using error or perf. on Dcv

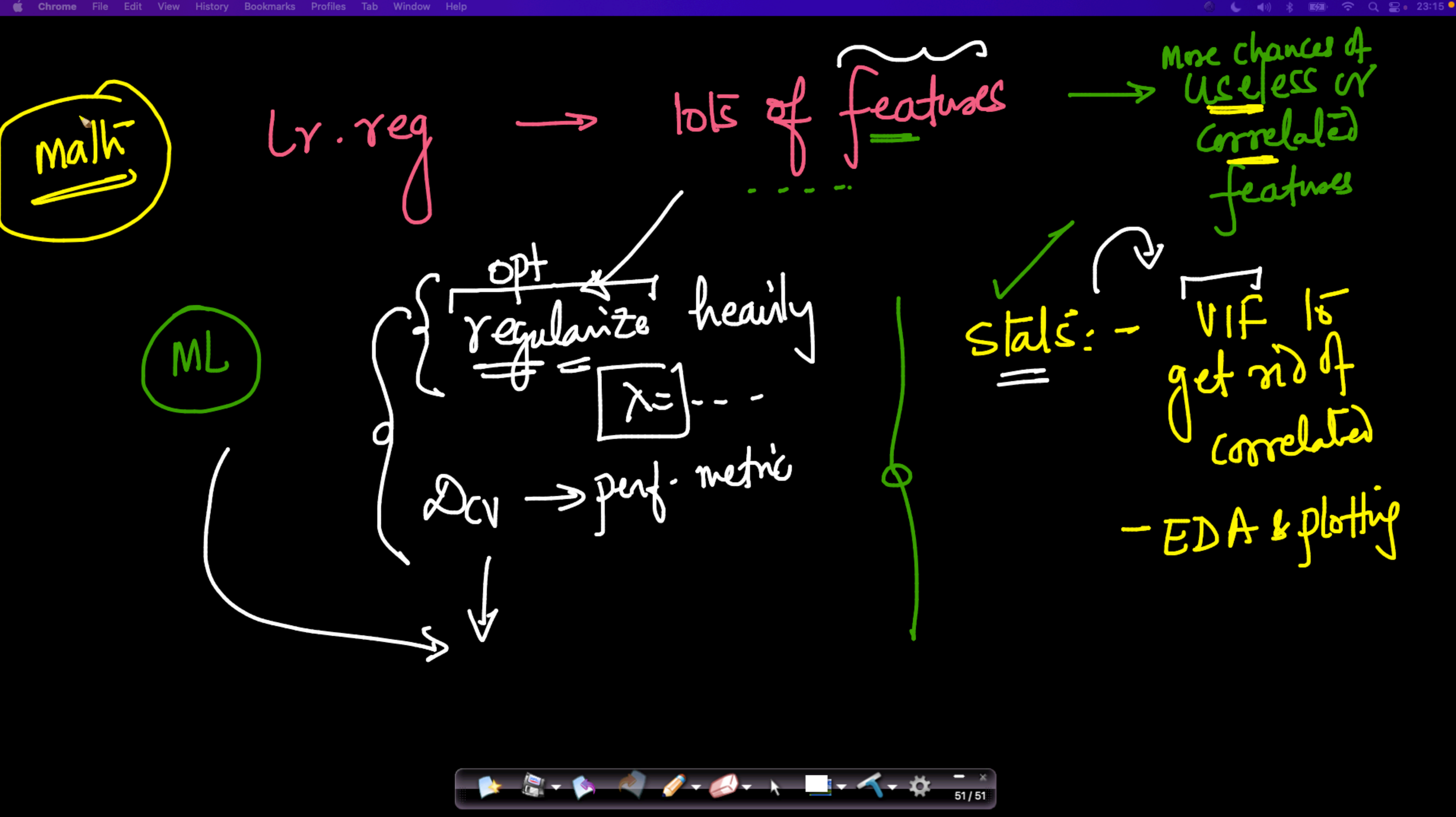
6 Report model-perf using Model.best on Test set

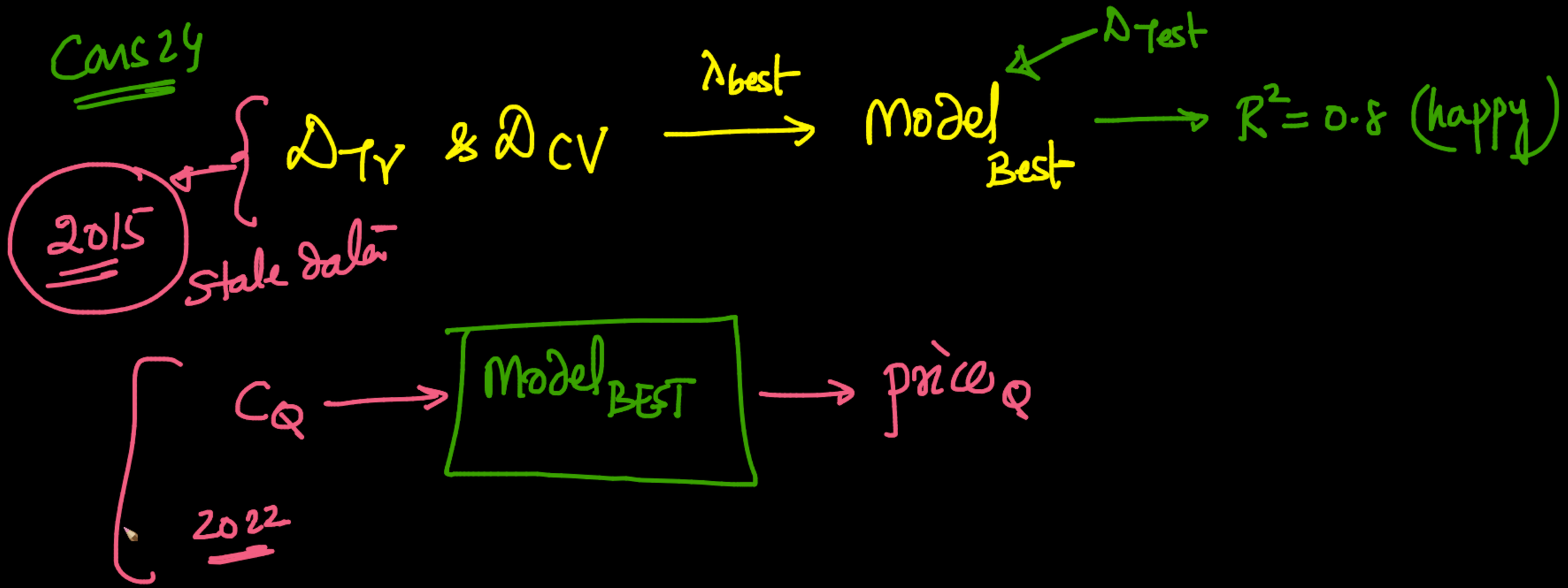


Leakage

↳ Some info about the leaks into your training process

mj & sj → Dtr → model params


Chrome File Edit View History Bookmarks Profiles Tab Window Help
23:15
Math
Lr.reg
lots of features
More chances of useless or correlated features
ML
opt
regularize heavily
λ= ---
2CV → perf. metric
Stats:-
VIF to get rid of correlated
- EDA & plotting
51 / 51



✓ {- retrain my model every nmlh- on recent data-

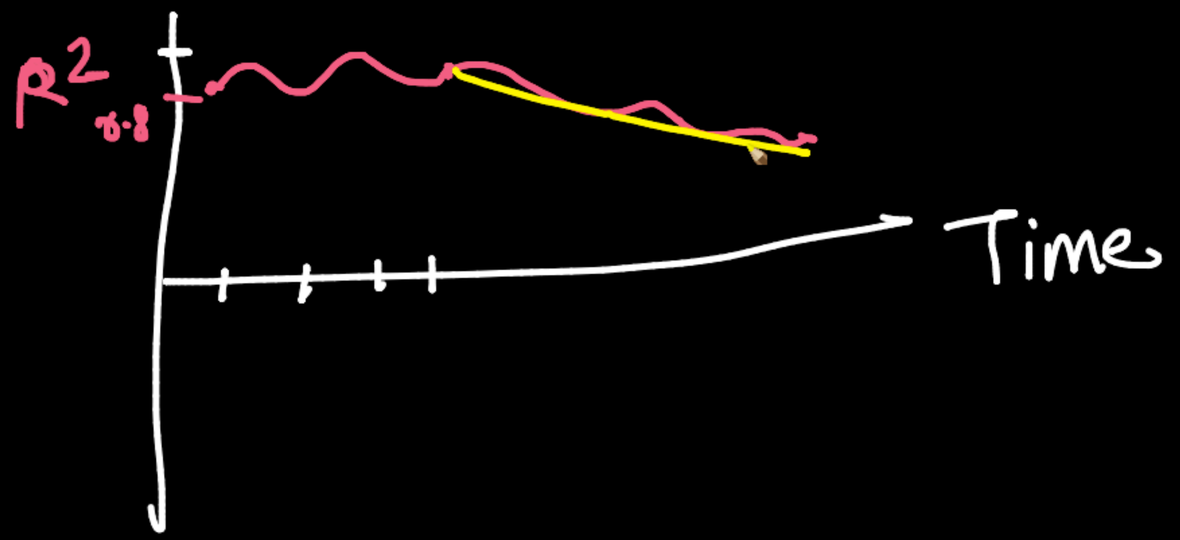
-

model is deteriorating over time

$D_{Test} \rightarrow R^2 = 0.8$

pred. price − act price

$\hat{y}_i$ y_i



d-features

Model Best

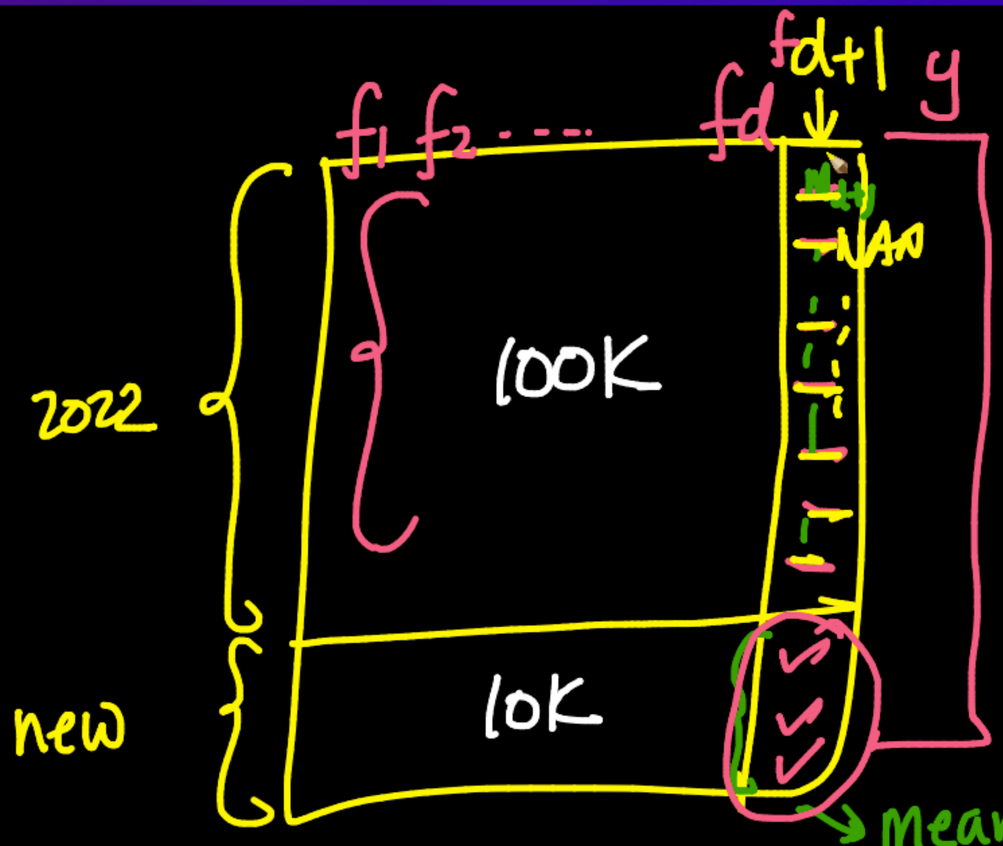
2022

d+1 st features

- VIF of d+1st feat with exisiting d-feat
- retain with all d+1st features

↓

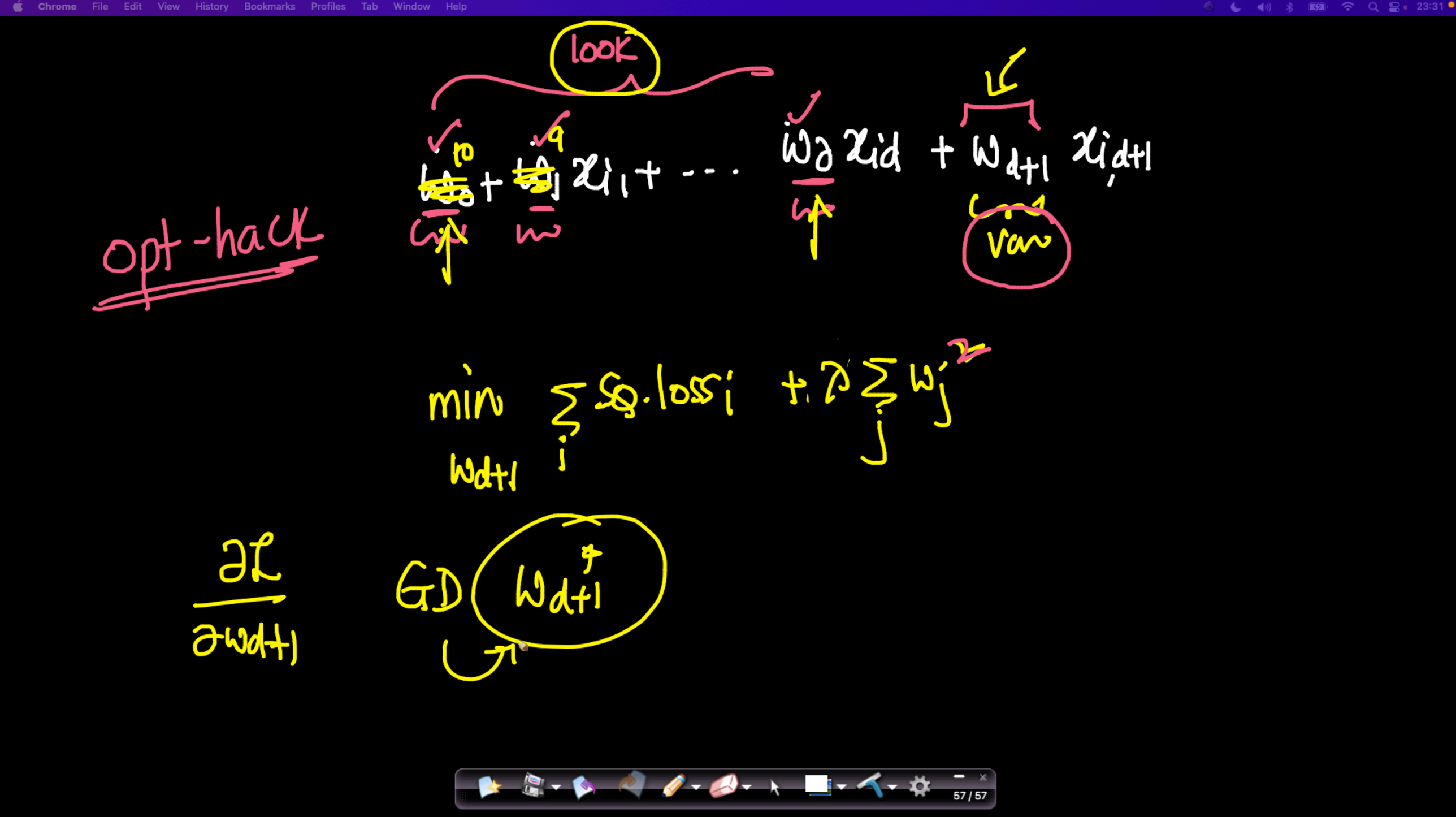
model perf improve



Model $_{\text{Best}}^{d}$

→ mean-value imputation

$w_1 \; \cdots \; w_d$


Chrome File Edit View History Bookmarks Profiles Tab Window Help
look
opt-hack
$w_0 + w_1 x_{i1} + \cdots w_d x_{id} + w_{d+1} x_{i,d+1}$
vary
$\min_{w_{d+1}} \sum_i sq.loss_i + \lambda \sum_j w_j^2$
$\frac{\partial L}{\partial w_{d+1}}$
GD
w_{d+1}
57 / 57

{ reuse & think deeply }

↓

- Code (trivial)
- logistic Regr...

weights

Test-error

100

$\lambda = 10$

8 non-zero params = 2 Zero pars

$\lambda = 20$

fewer

3 non-zero params = 7 zero pars

~~3~~ 120

$w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_{10} x_{i10}$

IoT slow

Fewer non-zero feat ⇒ few comp

- Computational
- Battery
- Memory
- Interpretability (Simpler) / Explainability