

Μηχανική Μάθηση για Προγνωστική Αναλυτική Δεδομένων

Fundamentals of Machine Learning for Predictive Data
Analytics

© John D. Kelleher and Brian Mac Namee and Aoife D'Arcy
Machine Learning in Business: An Introduction to the World
of Data Science.

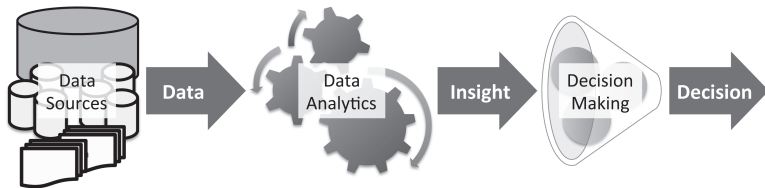
© John Hull

Αθανάσιος Σάκκας, ΟΠΑ

- 1 Τι είναι η Προγνωστική Αναλυτική Δεδομένων;
- 2 Τι είναι η Μηχανική Μάθηση;
- 3 Πώς Λειτουργεί η Μηχανική Μάθηση;
- 4 Επαγωγική Προκατάληψη έναντι Δειγματικής Προκατάληψης
- 5 Τι Μπορεί να Πάει Στραβά στη MM;
- 6 Σύνοψη
- 7 Εμπειρική Εφαρμογή
- 8 Παράδειγμα: Salary vs Age

Τι είναι η Προγνωστική Αναλυτική Δεδομένων;

- Η Προγνωστική Αναλυτική Δεδομένων περιλαμβάνει τις επιχειρησιακές και τις διαδικασίες δεδομένων, καθώς και τα υπολογιστικά μοντέλα που επιτρέπουν σε μια επιχείρηση να λαμβάνει **αποφάσεις βασισμένες σε δεδομένα**.



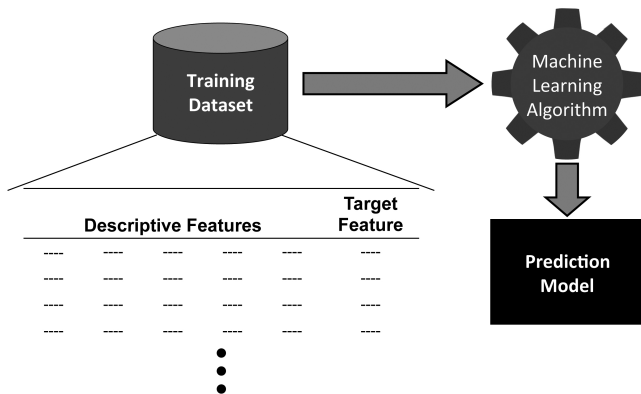
Σχήμα: Η προγνωστική αναλυτική δεδομένων μετακινείται από τα **δεδομένα** στις **γνώσεις/ευρήματα** και έπειτα στις **αποφάσεις**.

Παραδείγματα Εφαρμογών:

- Πρόβλεψη Τιμής
- Ανίχνευση Απάτης
- Πρόβλεψη Δοσολογίας
- Αξιολόγηση Κινδύνου
- Μοντελοποίηση Πιθανότητας Απόκρισης (Propensity modelling)
- Διάγνωση
- Ταξινόμηση Εγγράφων
- ...

Τι είναι η Μηχανική Μάθηση;

- Οι τεχνικές (επιβλεπόμενης) Μηχανικής Μάθησης μαθαίνουν αυτόματα ένα μοντέλο της σχέσης ανάμεσα σε ένα σύνολο **περιγραφικών χαρακτηριστικών** και σε ένα **χαρακτηριστικό-στόχο** από ένα σύνολο ιστορικών παραδειγμάτων.



Σχήμα: Χρήση της μηχανικής μάθησης για την επαγωγή ενός μοντέλου πρόβλεψης από ένα σύνολο δεδομένων εκπαίδευσης.



Σχήμα: Χρήση του μοντέλου για την παραγωγή προβλέψεων για νέα (ερωτηματικά) στιγμιότυπα.

ID	Occupation	Age	Loan-Salary	
			Ratio	Outcome
1	industrial	34	2.96	repaid
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repaid
7	professional	37	1.50	repaid
8	professional	40	1.93	repaid
9	industrial	33	5.25	default
10	industrial	32	4.15	default

- Ποια είναι η σχέση ανάμεσα στα **περιγραφικά χαρακτηριστικά** (Occupation, Age, Loan-Salary Ratio) και στο **χαρακτηριστικό-στόχο** (Outcome);

```

if Loan-Salary Ratio > 3 then
    Outcome='default'
else
    Outcome='repay'
end if
    
```

```

if Loan-Salary Ratio > 3 then
    Outcome='default'
else
    Outcome='repay'
end if
    
```

- Αυτό είναι ένα παράδειγμα **μοντέλου πρόβλεψης**

```

if Loan-Salary Ratio > 3 then
    Outcome='default'
else
    Outcome='repay'
end if
    
```

- Αυτό είναι ένα παράδειγμα **μοντέλου πρόβλεψης**
- Αυτό είναι επίσης ένα παράδειγμα **συνεπούς** μοντέλου πρόβλεψης

```
if Loan-Salary Ratio > 3 then
  Outcome='default'
else
  Outcome='repay'
end if
```

- Αυτό είναι ένα παράδειγμα **μοντέλου πρόβλεψης**
- Αυτό είναι επίσης ένα παράδειγμα **συνεπούς** μοντέλου πρόβλεψης
- Παρατηρήστε ότι αυτό το μοντέλο δεν χρησιμοποιεί όλα τα χαρακτηριστικά και ότι το χαρακτηριστικό που χρησιμοποιεί είναι παράγωγο (σε αυτή την περίπτωση ένας λόγος): ο **σχεδιασμός χαρακτηριστικών** και η **επιλογή χαρακτηριστικών** είναι δύο σημαντικά θέματα στα οποία θα επιστρέφουμε ξανά και ξανά.

- Ποια είναι η σχέση ανάμεσα στα **περιγραφικά χαρακτηριστικά** και στο **χαρακτηριστικό-στόχο** (Outcome) στο ακόλουθο σύνολο δεδομένων;

ID	Amount	Salary	Loan-Salary	Age	Occupation	House	Type	Outcome
			Ratio					
1	245,100	66,400	3.69	44	industrial	farm	stb	repaid
2	90,600	75,300	1.2	41	industrial	farm	stb	repaid
3	195,600	52,100	3.75	37	industrial	farm	ftb	default
4	157,800	67,600	2.33	44	industrial	apartment	ftb	repaid
5	150,800	35,800	4.21	39	professional	apartment	stb	default
6	133,000	45,300	2.94	29	industrial	farm	ftb	default
7	193,100	73,200	2.64	38	professional	house	ftb	repaid
8	215,000	77,600	2.77	17	professional	farm	ftb	repaid
9	83,000	62,500	1.33	30	professional	house	ftb	repaid
10	186,100	49,200	3.78	30	industrial	house	ftb	default
11	161,500	53,300	3.03	28	professional	apartment	stb	repaid
12	157,400	63,900	2.46	30	professional	farm	stb	repaid
13	210,000	54,200	3.87	43	professional	apartment	ftb	repaid
14	209,700	53,000	3.96	39	industrial	farm	ftb	default
15	143,200	65,300	2.19	32	industrial	apartment	ftb	default
16	203,000	64,400	3.15	44	industrial	farm	ftb	repaid
17	247,800	63,800	3.88	46	industrial	house	stb	repaid
18	162,700	77,400	2.1	37	professional	house	ftb	repaid
19	213,300	61,100	3.49	21	industrial	apartment	ftb	default
20	284,100	32,300	8.8	51	industrial	farm	ftb	default
21	154,000	48,900	3.15	49	professional	house	stb	repaid
22	112,800	79,700	1.42	41	professional	house	ftb	repaid
23	252,000	59,700	4.22	27	professional	house	stb	default
24	175,200	39,900	4.39	37	professional	apartment	stb	default
25	149,700	58,600	2.55	35	industrial	farm	stb	default

```

if Loan-Salary Ratio < 1.5 then
    Outcome='repay'
else if Loan-Salary Ratio > 4 then
    Outcome='default'
else if Age < 40 and Occupation ='industrial' then
    Outcome='default'
else
    Outcome='repay'
end if
    
```

```
if Loan-Salary Ratio < 1.5 then
  Outcome='repay'
else if Loan-Salary Ratio > 4 then
  Outcome='default'
else if Age < 40 and Occupation = 'industrial' then
  Outcome='default'
else
  Outcome='repay'
end if
```

- Η πραγματική αξία της μηχανικής μάθησης γίνεται εμφανής σε περιπτώσεις σαν κι αυτή, όταν θέλουμε να κατασκευάσουμε μοντέλα πρόβλεψης από μεγάλα σύνολα δεδομένων με πολλά χαρακτηριστικά.

Πώς Λειτουργεί η Μηχανική Μάθηση;

- Οι αλγόριθμοι μηχανικής μάθησης λειτουργούν αναζητώντας, μέσα σε ένα σύνολο πιθανών μοντέλων πρόβλεψης, το μοντέλο που αποτυπώνει καλύτερα τη σχέση μεταξύ των περιγραφικών χαρακτηριστικών και του χαρακτηριστικού-στόχου.

- Οι αλγόριθμοι μηχανικής μάθησης λειτουργούν αναζητώντας, μέσα σε ένα σύνολο πιθανών μοντέλων πρόβλεψης, το μοντέλο που αποτυπώνει καλύτερα τη σχέση μεταξύ των περιγραφικών χαρακτηριστικών και του χαρακτηριστικού-στόχου.
- Ένα προφανές κριτήριο αναζήτησης είναι να αναζητούμε μοντέλα που είναι **συνεπή** με τα δεδομένα.

- Οι αλγόριθμοι μηχανικής μάθησης λειτουργούν αναζητώντας, μέσα σε ένα σύνολο πιθανών μοντέλων πρόβλεψης, το μοντέλο που αποτυπώνει καλύτερα τη σχέση μεταξύ των περιγραφικών χαρακτηριστικών και του χαρακτηριστικού-στόχου.
- Ένα προφανές κριτήριο αναζήτησης είναι να αναζητούμε μοντέλα που είναι **συνεπή** με τα δεδομένα.
- Ωστόσο, επειδή το σύνολο εκπαίδευσης είναι μόνο ένα δείγμα, η μηχανική μάθηση είναι ένα **κακώς τεθειμένο** (ill-posed) πρόβλημα.

Πίνακας: Ένα απλό σύνολο δεδομένων λιανικής

ID	Bby	Alc	Org	Grp
1	no	no	no	couple
2	yes	no	yes	family
3	yes	yes	no	family
4	no	no	yes	couple
5	no	yes	yes	single

Πίνακας: Ένα πλήρες σύνολο πιθανών μοντέλων πρόβλεψης πριν γίνει διαθέσιμο οποιοδήποτε δεδομένο εκπαίδευσης.

Bby	Alc	Org	Grp	M ₁	M ₂	M ₃	M ₄	M ₅	...	M ₆ 561
no	no	no	?	couple	couple	single	couple	couple		couple
no	no	yes	?	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	?	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	?	couple	family	family	family	family		couple
yes	yes	no	?	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

Πίνακας: Ένα δείγμα από τα μοντέλα που είναι συνεπή με τα δεδομένα εκπαίδευσης

Bby	Alc	Org	Grp	M ₁	M ₂	M ₃	M ₄	M ₅	...	M ₆ 501
no	no	no	couple	couple	couple	single	couple	couple		couple
no	no	yes	couple	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	single	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	family	couple	family	family	family	family		couple
yes	yes	no	family	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

Πίνακας: Ένα δείγμα από τα μοντέλα που είναι συνεπή με τα δεδομένα εκπαίδευσης

Bby	Alc	Org	Grp	M ₁	M ₂	M ₃	M ₄	M ₅	...	M ₆ 501
no	no	no	couple	couple	couple	single	couple	couple		couple
no	no	yes	couple	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	single	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	family	couple	family	family	family	family		couple
yes	yes	no	family	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

- Παρατηρήστε ότι απομένουν περισσότερα από ένα υποψήφια μοντέλα! Επειδή δεν μπορεί να βρεθεί ένα μοναδικό συνεπές μοντέλο μόνο από ένα **δειγματικό** σύνολο εκπαίδευσης, η MM είναι **κακώς τεθειμένη**.

- Η συνέπεια $\approx$ απομνημόνευση του συνόλου δεδομένων.
- Η συνέπεια με θόρυβο στα δεδομένα δεν είναι επιθυμητή.
- Στόχος: ένα μοντέλο που γενικεύει πέρα από το σύνολο δεδομένων και δεν επηρεάζεται από τον θόρυβο.
- Άρα, ποιο κριτήριο πρέπει να χρησιμοποιήσουμε για να επιλέξουμε μεταξύ μοντέλων;

- **Επαγωγική προκατάληψη (Inductive bias):** το σύνολο των υποθέσεων που ορίζουν τα κριτήρια επιλογής μοντέλου ενός αλγορίθμου MM.
- Υπάρχουν δύο τύποι προκατάληψης που μπορούμε να χρησιμοποιήσουμε:
 - 1 προκατάληψη περιορισμού (restriction bias)
 - 2 προκατάληψη προτίμησης (preference bias)
- Η επαγωγική προκατάληψη είναι απαραίτητη για τη μάθηση (πέρα από το σύνολο δεδομένων).

Πώς λειτουργεί η MM (Σύνοψη)

- Οι αλγόριθμοι MM λειτουργούν αναζητώντας μέσα σε σύνολα πιθανών μοντέλων.
- Υπάρχουν δύο πηγές πληροφορίας που καθοδηγούν αυτή την αναζήτηση:
 - 1 τα δεδομένα εκπαίδευσης,
 - 2 η επαγωγική προκατάληψη του αλγορίθμου.

Επαγωγική Προκατάληψη έναντι Δειγματικής Προκατάληψης

- Η επαγωγική προκατάληψη είναι απαραίτητη για τη μηχανική μάθηση και, κατά μία έννοια, βασικός στόχος του αναλυτή δεδομένων είναι να βρει τη σωστή επαγωγική προκατάληψη.
- Η επαγωγική προκατάληψη δεν είναι ο μόνος τύπος προκατάληψης που επηρεάζει τη μηχανική μάθηση· ένας ιδιαίτερα σημαντικός τύπος που πρέπει να γνωρίζουμε είναι η **δειγματοληπτική προκατάληψη** (sampling bias).

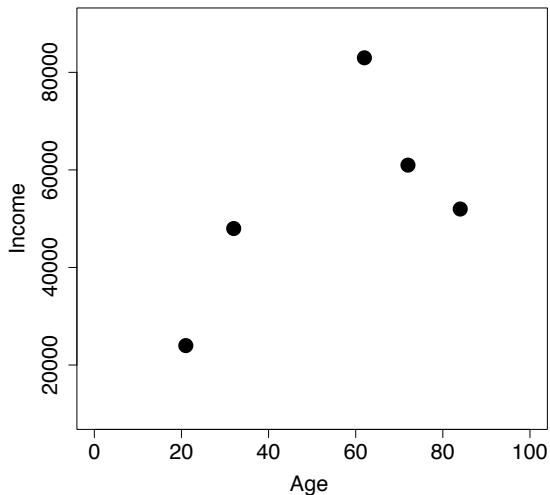
- Η δειγματοληπτική προκατάληψη προκύπτει όταν το δείγμα δεδομένων που χρησιμοποιείται σε μια διαδικασία βασισμένη σε δεδομένα συλλέγεται με τρόπο ώστε να μην είναι αντιπροσωπευτικό του πληθυσμού που υποτίθεται ότι εκπροσωπεί.
- Αν ένα δείγμα δεν είναι αντιπροσωπευτικό του πληθυσμού, τότε τα συμπεράσματα που βασίζονται σε αυτό δεν θα γενικεύονται στον ευρύτερο πληθυσμό.
- Η δειγματοληπτική προκατάληψη είναι κάτι που ο αναλυτής δεδομένων πρέπει προληπτικά να προσπαθεί να αφαιρεί από τα δεδομένα σε κάθε έργο αναλυτικής δεδομένων.

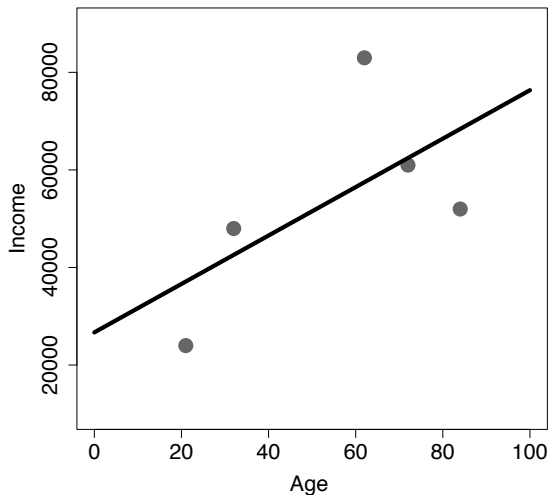
Τι Μπορεί να Πάει Στραβά στη
MM;

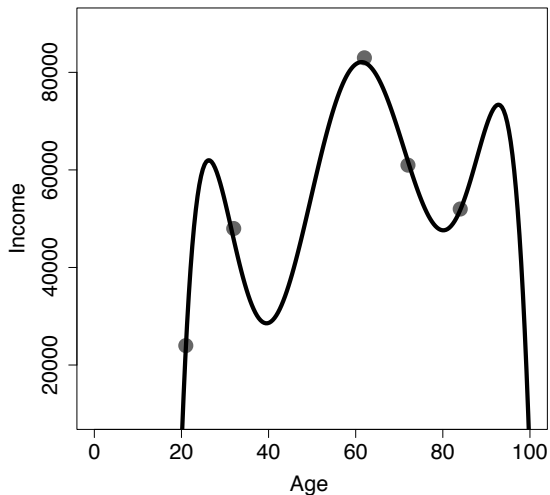
- No free lunch!
- Τι συμβαίνει αν επιλέξουμε λάθος επαγωγική προκατάληψη:
 - 1 **υποπροσαρμογή** (underfitting)
 - 2 **υπερπροσαρμογή** (overfitting)

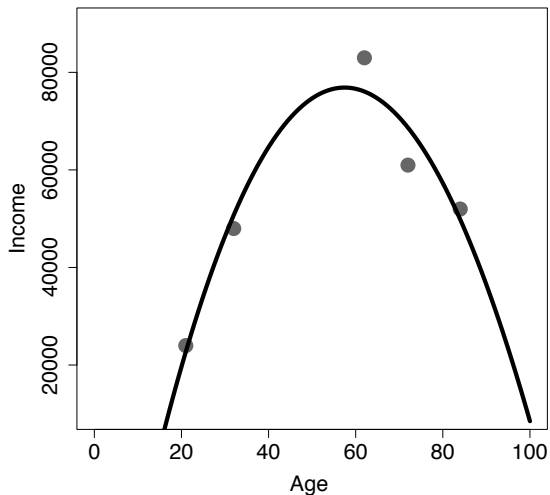
Πίνακας: Το σύνολο δεδομένων ηλικίας-εισοδήματος.

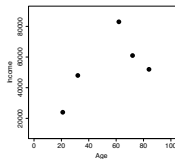
ID	Age	Income
1	21	24,000
2	32	48,000
3	62	83,000
4	72	61,000
5	84	52,000



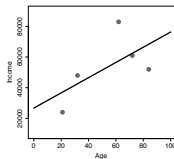




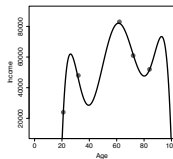




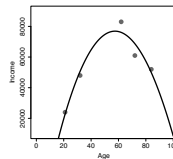
(α□) Σύνολο δε-
δομένων



(β□) Υποπρο-
σαρμογή



(γ□) Υπερπρο-
σαρμογή



(δ□) Ακριβώς
σωστό

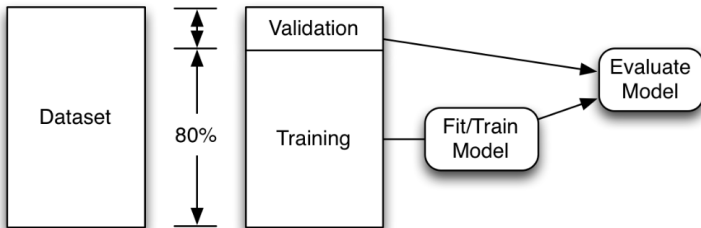
Σχήμα: Εξισορρόπηση μεταξύ υπερπροσαρμογής και υποπροσαρμογής όταν προσπαθούμε να προβλέψουμε την ηλικία από το εισόδημα.

Σύνοψη

- Οι τεχνικές Μηχανικής Μάθησης μαθαίνουν αυτόματα τη σχέση ανάμεσα σε ένα σύνολο **περιγραφικών χαρακτηριστικών** και σε ένα **χαρακτηριστικό-στόχο** από ένα σύνολο ιστορικών παραδειγμάτων.
- Η Μηχανική Μάθηση είναι ένα **κακώς τεθειμένο** (ill-posed) πρόβλημα:
 - 1 **γενίκευση** (generalize),
 - 2 **επαγωγική προκατάληψη** (inductive bias),
 - 3 **υποπροσαρμογή** (underfitting),
 - 4 **υπερπροσαρμογή** (overfitting).
- Η επίτευξη της σωστής ισορροπίας μεταξύ πολυπλοκότητας και απλότητας (μεταξύ υπερπροσαρμογής και υποπροσαρμογής) είναι το δυσκολότερο μέρος της μηχανικής μάθησης.

Εμπειρική Εφαρμογή

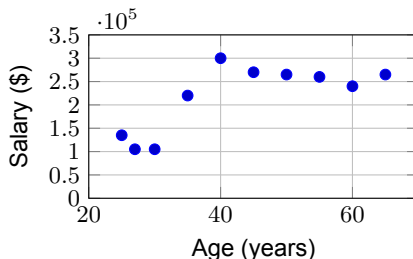
ML Good Practice (2)



Παράδειγμα: Μισθός σε συνάρτηση με την ηλικία

- Μισθός σε συνάρτηση με την ηλικία για ένα συγκεκριμένο επάγγελμα σε μια συγκεκριμένη περιοχή.
- (Στο αρχικό υλικό: *Salary vs. Age Example.xlsx*)

Age	Salary (\$)
25	135,000
55	260,000
27	105,000
35	220,000
60	240,000
65	265,000
45	270,000
40	300,000
50	265,000
30	105,000

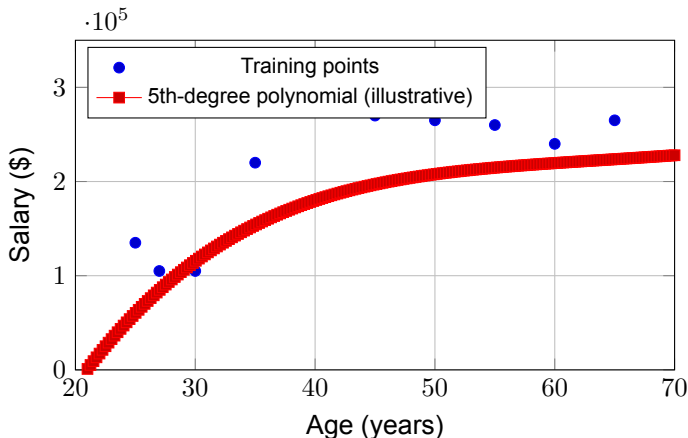


Scatter plot για το training set

A Good Fit (Y = Salary, X = Age)

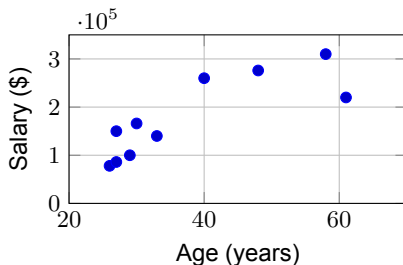
$$Y = a + b_1X + b_2X^2 + b_3X^3 + b_4X^4 + b_5X^5$$

- To root mean squared error (rmse) για το training dataset είναι **\$12,902**.



Ένα Out-of-Sample Validation Set

Age	Salary (\$)
30	166,000
26	78,000
58	310,000
29	100,000
40	260,000
27	150,000
33	140,000
61	220,000
27	86,000
48	276,000

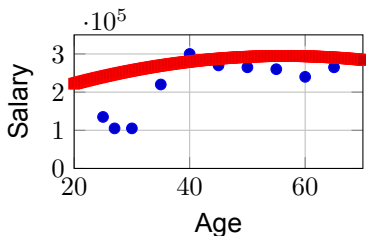


Scatter plot για το Validation Set

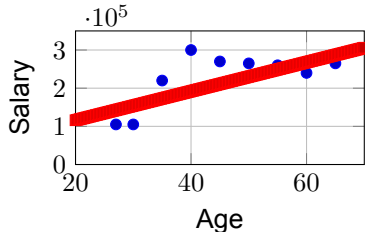
- $\text{rmse}(\text{training}) = \$12,902$.
- $\text{rmse}(\text{validation}) = \$38,794$ (πολύ υψηλότερο).
- Συμπέρασμα: το μοντέλο **overfits** και δεν γενικεύεται καλά.

Quadratic Model & Linear Model

$$Y = a + b_1X + b_2X^2$$



$$Y = a + b_1X$$



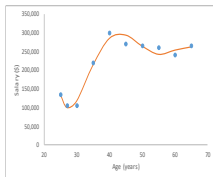
- Quadratic: rmse training = **\$32,932**, rmse validation = **\$33,554**.
- Linear: rmse training = **\$49,731**, rmse validation = **\$49,990**.

Σύνοψη Αποτελεσμάτων: RMSE

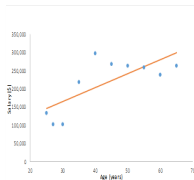
	Polynomial (deg 5)	Quadratic	Linear
Training set	12,902	32,932	49,731
Validation set	38,794	33,554	49,990

- Linear model: **underfits**. Το rmse του training set είναι όσο μεγάλο είναι και το rmse του validation set
- 5th degree polynomial: **overfits**. το rmse του training set είναι πολύ μικρό ενώ, το rmse του validation set είναι πολύ μεγαλύτερο
- Το quadratic μοντέλο είναι το πιο ακριβές, έχει το μικρότερο rmse στο rmse από τα υπόλοιπα μοντέλα και δεν έχει μεγάλη διαφορά από το rmse του training set Αλλά πόσο ακριβές είναι το quadratic μοντέλο. Η απάντηση στις επόμενες διαφάνειες.

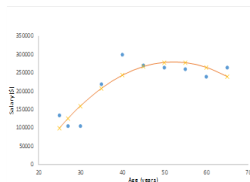
Overfitting / Underfitting (Διαισθητικά)



Overfitting



Underfitting



Best

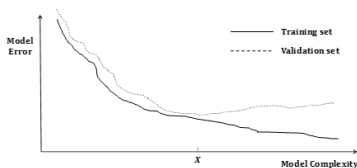
Test set: Τελικός έλεγχος (out-of-sample)

Πρέπει να υπολογίσουμε το rmse του test dataset.

Age	Salary (\$)	Predicted (\$)	Error (\$)
26	110,000	113,172	-3,172
52	278,000	279,589	-1,589
38	314,000	232,852	+83,148
60	302,000	264,620	+37,380
64	261,000	245,457	+15,543
41	227,000	249,325	-22,325
34	200,000	199,411	+589
46	233,000	270,380	-37,380
57	311,000	273,883	-37,117
55	298,000	277,625	+20,375

- To SD of error (rmse) είναι **\$34,273**.
- Rule of thumb: αυξάνουμε πολυπλοκότητα έως ότου το μοντέλο δεν γενικεύεται πλέον καλά στο out-of-sample set.

Typical Pattern of Errors (Training vs Validation)



- **Bias-variance trade-off**
- Bias: σφάλμα από underfitting.
- Variance: σφάλμα από overfitting.