Lec5 - Regression Losses & Metrics

Monday, 1 July 2024 16:13

תזכורת -

mse (בונ' מחיר) loss : פונקציה שמחפש מינימום שלה למציאת פרמטרים של מודל

mse (ביצועים) metrics : מדד לביצועים (לפעמים זה loss, אבל לא חייב)

מטרה: loss/metrics עבור חיזוי

מהירות

שילוב לחישוב מינימום של פונ' :

② * מבוסס נגזרת ראשונה
דוגמה gradient descent

① * נגזרת ראשונה + שניה

③ * ללא נגזרת

הקדמה: עבור פונ' loss

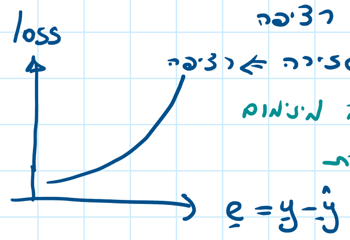
* מרחק בין $\underline{y}$ "אמיתי" ל $\underline{\hat{y}}$ שהוא מוצא של מודל

* עד היום MSE

* פונקציה רציפה

* גזירה ← רציפה

רוב "ממ" חישוב מינימום מבוססים נגזרת



* Convex = נקודת מינימום גלובלי יחיד

General Losses and Metrics

Per-sample error, $e_i = y_i - \hat{y}_i, i = 1, \dots, M$.
Vector notation, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.

סדר לא משנה

מספר שורות (=דוגמאות) בבסיס נתונים

Metrix

MSE (L_2-loss)

$$J(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{M}\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \frac{1}{M}\mathbf{e}^T\mathbf{e}$$

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

Loss

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2M}\sum_{i=1}^{M} e_i^2$$

בשביל להתכסס עם נגזרת

* לא חייבים ממוצע לרגול כי לא משפיע על נקודת מינימום

* כן חייבים עם regularization וכו'

תכונות mse

* רציף

* גזיר

* מינימום גלובלי יחיד

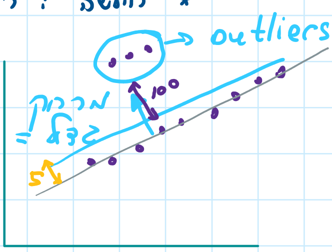
* נפוץ מאוד - נתמך ע"י חבילות תוכנה

* נגזרת יחסית לגודל השגיאה
- חשוב ל-GD

* לפעמים יש פתרון אנליטי

* "חולשה" : רגישות ל - outliers

נקודת חריגות



= השפעה גדולה מאוד (ריבוע המרחק) על MSE

דוגמה מספרית :

e	MSE
10	100
100	10,000

root-MSE

שורש של MSE

$$J(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{\sqrt{M}} \|\mathbf{y} - \hat{\mathbf{y}}\|$$
$$= \sqrt{\frac{1}{M} \sum_{i=1}^{M} (y_i - \hat{y})^2} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} e_i^2}$$

* שווה ל- MSE בנגזרת ⇐ בד"כ עובדים
עם MSE כ- loss
RMSE → metric ביחידות זהות ל- y
≈ סט"ת תקן

Mean absolute error (MAE)

MAE used both as loss and metric.

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{M} \sum_{i=1}^{M} |y_i - \hat{y}| = \frac{1}{M} \sum_{i=1}^{M} |e_i|$$

* שימושי מאוד
* בתחום רגיש ל- outliers
* נגזרת בלתי תלויה בשגיאה
← נגזרת מאוד גבוהה (קבועה)
גם כשהשגיאה מאוד נמוכה
עבור MSE: נגזרת יורדת בהדרגה
עם הקטנת השגיאה
עשוי לגרום בעיית התכנסות
בעיקר בעזירה
+ מקדם α משתנה
עבור GD

* לא נגזרת עבור $y = \hat{y}$
← לא השפעה משמעותית

Huber loss

Goal: Hybrid between MAE and MSE

$$\mathcal{L}(e_i) = \begin{cases} \frac{1}{2} e_i^2 & |e_i| \le \delta \\ \delta \left(|e_i| - \frac{1}{2}\delta \right) & \text{otherwise} \end{cases}$$

~ MSE
~ MAE

* רציף
* גזיר
* נקודת מינימום גלובלי
* חסרון: δ - hyper-parameter
רציף ← דורש חיפוש
ערך אופטימלי

Log-cosh loss

Goal: Hybrid between MAE and MSE without hyper-parameters.

$$\mathcal{L}(e_i) = \log \cosh e_i$$

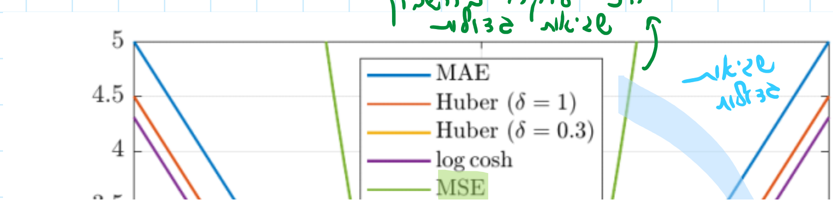
בקירוב:

$$L(e_i) \approx \begin{cases} \frac{e_i^2}{2} & e_i \text{ small} \\ |e_i| - \log(2) & e_i \text{ large} \end{cases}$$

MSE
MAE

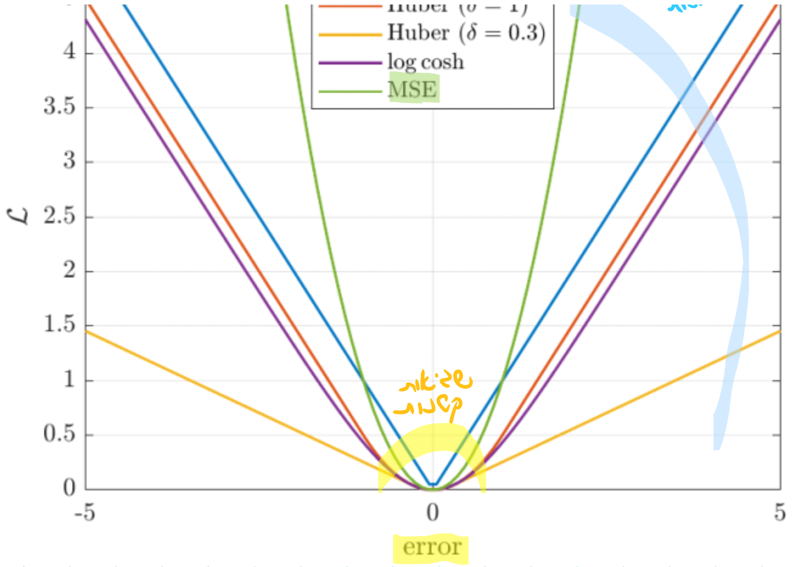
* גזיר $\frac{\partial}{\partial e_i} L(e_i) = \tanh e_i$
* אין היפר-פרמטרים
* חסרון: בפועל לא עוזר,
כי צריך להגדיל שגיאה
בתחום מתאים

הכי לוקח בחשבון
שגיאות גדולות



שגיאות גדולות

סיכום ביניים:
עיקר ההבדל
זה בהתמודדות



עיקר ההבדל
זה התמודדות
עם Outliers

Mean Squared Logarithmic Error (MLSE)

מטרה: הבדלים משמעותיים בערכי y
⇐ תווך דינמי גדול

$$J(y_i, \hat{y}_i) = \frac{1}{M}\sum_{i=1}^{M}\left(\log(y_i+1) - \log(\hat{y}_i+1)\right)^2$$
$$= \frac{1}{M}\sum_{i=1}^{M}\log\left(\frac{y_i+1}{\hat{y}_i+1}\right)^2$$

שומר על היחס
לא סימטרי
חסרון העיקרי

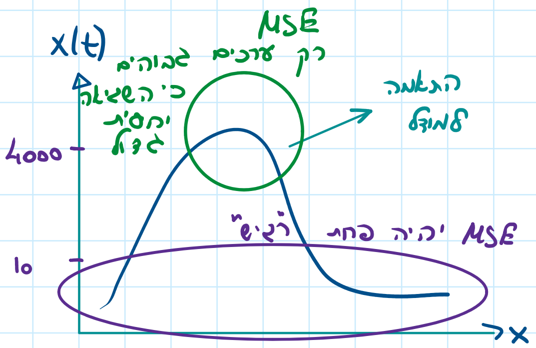
Table 6.1: Example of MLSE

True Values	Predicted Values	MSE Loss	MSLE Loss
40	30	100	0.078168
4000	3000	1000000	0.082714
20	10	100	**0.07886**
20	30	100	**0.02861**

* 1+ כדי לאפשר $y_i, \hat{y}_i = 0$

$\log(0) \to -\infty$

Root MLSE (RMLSE) * קיים גם בגרסה



Metrics

מדדי ביצועים מנורמלים
* משמשים להשוואה בין מודלים
* בד"כ בתחום בין 0 ל-1
* מובנים אנושים

Relative squared error (RSE)

MSE מנורמל

$$J(y_i, \hat{y}_i) = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = \frac{MSE}{\mathrm{Var}[\mathbf{y}]} = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2}$$

* הקטנת שונות השגיאה
* יותר קרוב ל-0 ⇐ יותר טוב

* עבור חיזוי נאיבי
חיזוי הממוצע
$\hat{y} = E[y]$
$\Rightarrow \quad MSE = Var[y]$

* הקטע שומר השגיאה
* יו' קרוב ל-0 ⇐ יו' טוב
* בד"כ קטן מ-1

מדעי הנדסה

R^2

מדעי החברה

$R^2 = 1 - RSE$

$R^2 \in [0,1]$

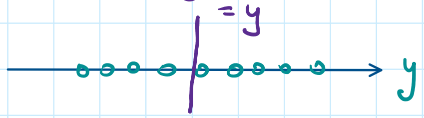
בנסיבות חריגות שלילי

יו' טוב

$\Rightarrow \quad MSE = Var[y]$

מטרה:

רוצים של y $\hat{y} = E[y] = \bar{y}$



$y = b$

$$MSE = \sum_i (y_i - b)^2$$

$$\frac{d}{db} mse = \cancel{2}\sum_i (y_i - b)\cancel{(-1)} = 0$$

$$\sum_{i=1}^{M} y_i - \underbrace{\sum_{i=1}^{M} b}_{M \cdot b} = 0$$

$$MSE = \sum_i (y_i - \bar{y})^2 \leftarrow b = \frac{1}{M}\sum_{i=1}^{M} y_i = \bar{y}$$
$$= Var[y]$$

* תוצאה שלא "נעזרת" במשתנים אחרים
⇐ כמו של מודל מבוסס כניסה נוספת
הנע מ-y להיות יו' טוב.

Relative absolute error (RAE)

Normalized MAE loss.

$$J(y_i, \hat{y}_i) = \frac{\sum_i |y_i - \hat{y}_i|}{\sum_i |y_i - \bar{y}|} = \frac{MAE}{\sum_i |y_i - \bar{y}|}$$

Closer to 0 is better.

True Values	Predicted Values	Absolute Error	MAPE
100	60	40	40%
20	60	40	**300%**

Mean Absolute Percentage Error (MAPE)

שגיאה יחסית באחוזים

Scaled error metric.

$$J = \frac{1}{M}\sum_{j=1}^{M} \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \qquad (6.$$

- Beware of small denominator!
- Can exceed 100%.

* בעיה: y_i קטן

משפחה של פונ' Loss

מטרה: לשלב מספר פרמטרים במודל

שיטה: סוג של "קנס" של loss עבור מספר פרמטרים גדול

בעיה: הרבה פרמטרים ⇐ overfit

סיבוכיות חישוב

לא בדיוק שיפור "אמיתי" של ביצועים

- N is the number of parameters,
- M is the number of data-points.

* כל השיטות: מבוססות MSE
* יש ביסוס תאורטי "עמוק"
* משמשות להשוואה של מודלים דומים בעל מספר פרמטרים שונה

Akaike's Final Prediction Error (FPE)

Akaike's Final Prediction Error (FPE) criterion provides a measure of model quality.

$$FPE = MSE\frac{1+N/M}{1-N/M} = MSE\frac{M+N}{M-N} \qquad (6.16)$$

MSE + נוסחה מבוססת מספר נקודות M
ומספר פרמטרים N
⇓ ומתצירה מספר

Akaike's Information Criterion Penalize the number of parameters,

$$\Delta AIC = 2N + M\ln(MSE) \qquad (6.17)$$

AICc is AIC with a correction for small sample sizes

$$AICc = AIC + 2N\frac{N+1}{M-N-1} \qquad (6.18)$$

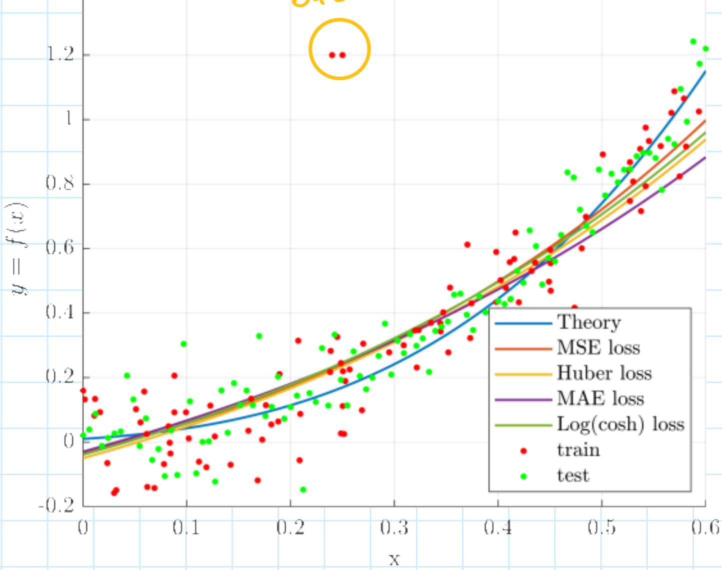
מספר המתקבל משמש להשוואה של מודלים
⇓
בוחרים מודל עם מספר מינימלי.

Bayesian Information Criterion (BIC)

$$BIC = M\ln(MSE) + N\ln(M)$$

דוגמה מספרית

outliers



דוגמה של מודל פולינומיאלי
* דורש אופטימיזציה של
– λ – פרמטר regular
– δ עבור loss Huber
עבור כל loss בנפרד
* כל שיטה: תוצאה קצת אחר

```
function loss = huber_loss(y,yh,delta)
    e = abs(y - yh);
    loss = zeros(size(e));
    idx = e<=delta;
    loss(idx) = e(idx).^2/2;
    loss(~idx) = delta*e(~idx) - delta^2/2;
    loss = mean(loss);
end
function loss = log_cosh_loss(y,yh)
    loss = mean(log(cosh(y-yh)));
end
function loss = mae_loss(y,yh)
    e = abs(y - yh);
    loss = mean(e);
end
```

$$\mathcal{L}(e_i) = \begin{cases} \frac{1}{2}e_i^2 & |e_i| \le \delta \\ \delta\left(|e_i| - \frac{1}{2}\delta\right) & \text{otherwise} \end{cases}$$