Machine Learning Course - CS-433

Cost Functions

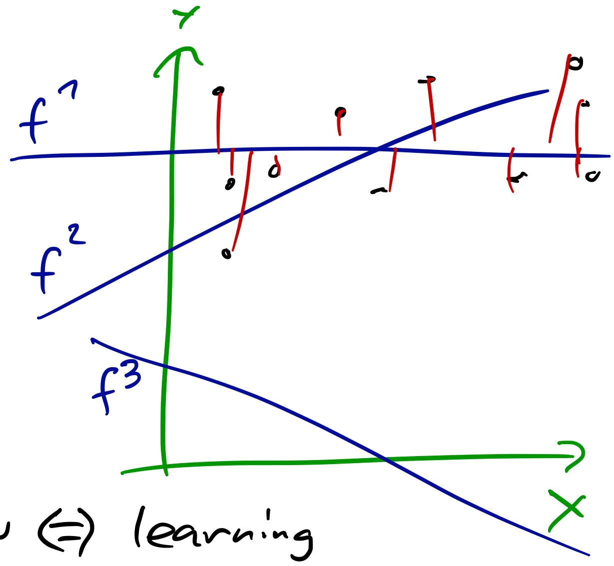
Sep 22, 2016

minor changes by Martin Jaggi 2016

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

Consider the following models.



1-parameter model: $y_n \approx w_0$

2-parameter model: $y_n \approx w_0 + w_1 x_{n1}$

How can we estimate (or guess) values of $\mathbf{w}$ given the data $\mathcal{D}$?

find w $\Leftrightarrow$ learning

What is a cost function?

A cost function (or energy, loss, training objective) is used to learn parameters that explain the data well. The cost function quantifies how well our model does - or in other words how costly our mistakes are.

Two desirable properties of cost functions

When the target y is real-valued, it is desirable that the cost is symmetric around 0, since both positive and negative errors should be penalized equally.

Also, our cost function should penalize "large" mistakes and "very-large" mistakes similarly.

Statistical vs computational trade-off

If we want better statistical properties, then we have to give-up good computational properties.

find true f

find f quickly

Mean Square Error (MSE)

MSE is one of the most popular cost functions.

$f = f_w$
$= x^T w$

$$\text{MSE}(\mathbf{w}) := \frac{1}{N}\sum_{n=1}^{N} \left[y_n - f(\mathbf{x}_n) \right]^2$$

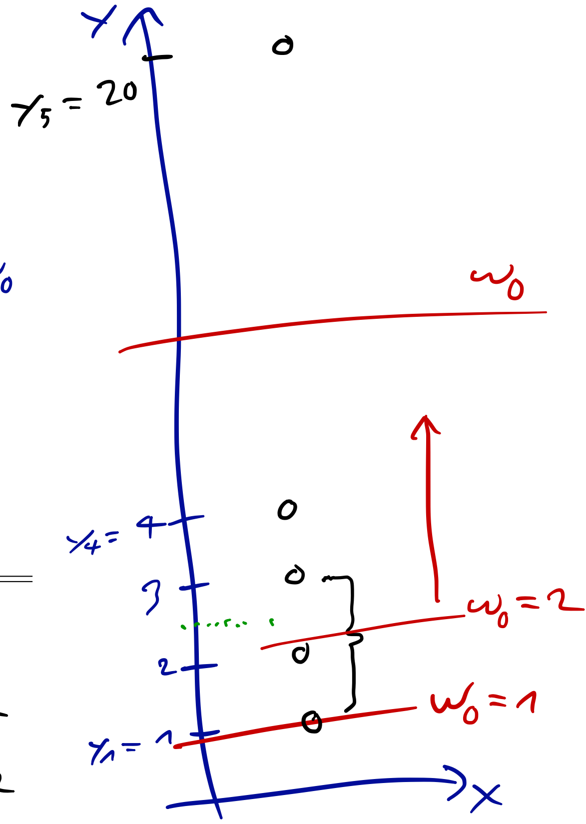
Does this cost function have both mentioned properties?

An exercise for MSE

Compute MSE for 1-param model:

$f(x) := w_0$

$$\mathcal{L}(w_0) := \frac{1}{N}\sum_{n=1}^{N=4 \text{ or } 5} \left[y_n - w_0 \right]^2$$



w_0	1	2	3	4	5	6	7
$y_1 =$ 1	0	1	2^2	3^2			6^2
$y_2 =$ 2	1^2	0	1	2^2			5^2
$y_3 =$ 3	2^2	1	0	1			4^2
$y_4 =$ 4	3^2	2^2	1	0			3^2
$N \cdot$ MSE	14	6	6	14	..	..	..
$y_5 = 20$	361	324	289	256	225	196	169
$N \cdot$ MSE						250	

→ $w_0 = 2$ or 3 is best

→ $w_0 = 6$ is best

Some help: $19^2 = 361, 18^2 = 324, 17^2 = 289, 16^2 = 256, 15^2 = 225, 14^2 = 196, 13^2 = 169$.

Outliers

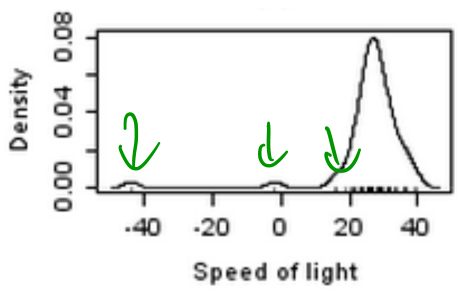
Outliers are data examples that are far away from most of the other examples. Unfortunately, they occur more often in reality than you would want them to!

MSE is not a good cost function when outliers are present.

Here is a real example on speed of light measurements (Gelman's book on Bayesian data analysis)

28	26	33	24	34	-44	27	16	40	-2
29	22	24	21	25	30	23	29	31	19
24	20	36	32	36	28	25	21	28	29
37	25	28	26	30	32	36	26	30	22
36	23	27	27	28	27	31	27	26	33
26	32	32	24	39	28	24	25	32	25
29	27	28	29	16	23				

(a) Original speed of light data done by Simon Newcomb.

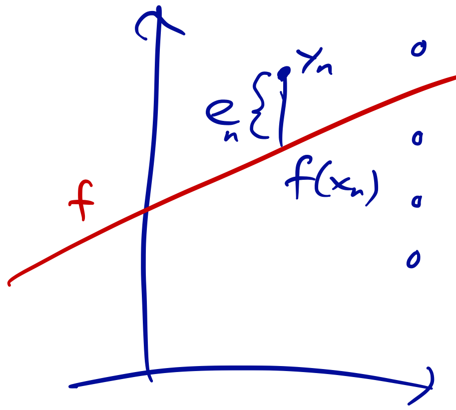


(b) Histogram showing outliers.

Handling outliers well is a desired *statistical* property.

Mean Absolute Error (MAE)

$f = f_w$



$$\text{MAE}(\mathbf{w}) := \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left|y_n - f(\mathbf{x}_n)\right|}_{e_n}$$

$e_n := |y_n - f(x_n)|$

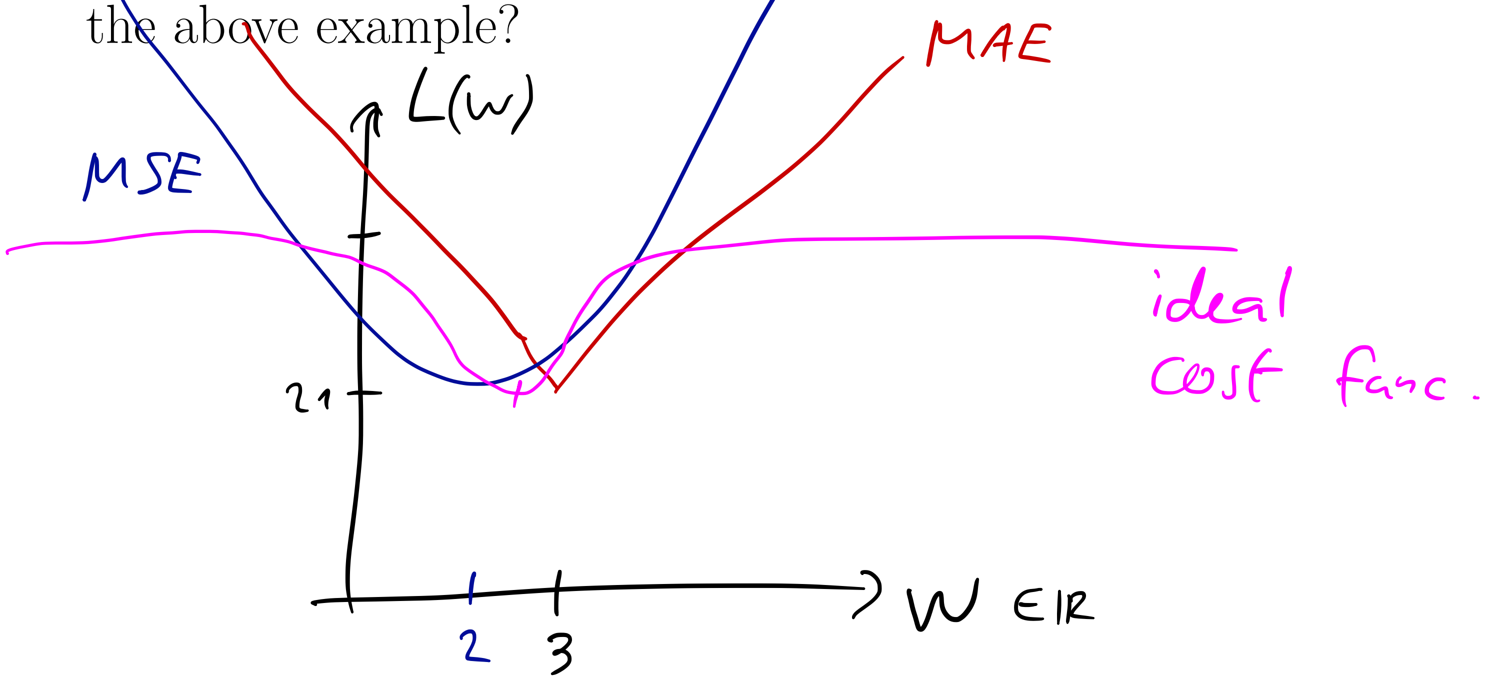
Repeat the exercise with MAE.

$w_0 =$	1	2	3	4	5	6	7
$y_1 =$ 1	0	1	2	3	4	5	6
2	1	0	1	2	3	4	5
3	2	1	0	1	2	3	4
$y_4 =$ 4	3	2	1	0	1	2	3
$N \cdot$ MAE	6	4	4	6	10	.	.
$y_5 = 20$	19	18	17	16	15	14	13
$N \cdot$ MAE	25	22	21	22	25	.	.

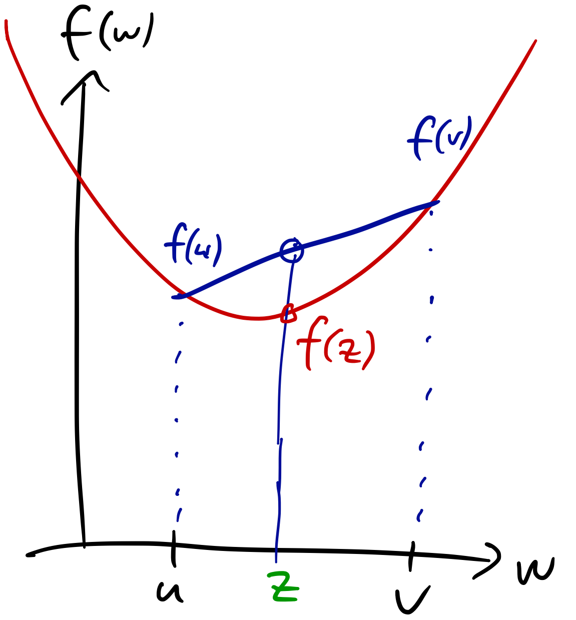
best model: $w_0 = 2, 3$

best m.: $w_0 = 3$

Can you draw MSE and MAE for the above example?



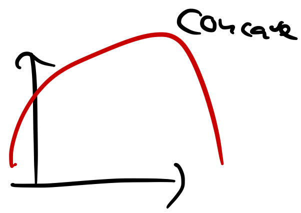
Convexity



Roughly, a function is convex iff a line joining two points never intersects with the function anywhere else.

A function $f(\mathbf{u})$ with $\mathbf{u} \in \mathcal{X}$ is convex, if for any $\mathbf{u}, \mathbf{v} \in \mathcal{X}$ and for any $0 \leq \lambda \leq 1$, we have:

$$f(\underbrace{\lambda \mathbf{u} + (1-\lambda)\mathbf{v}}_{z}) \leq \lambda f(\mathbf{u}) + (1-\lambda) f(\mathbf{v})$$



$<$

A function is strictly convex if the inequality is strict. for $0 < \lambda < 1$

Importance of convexity

A convex function has only one global minimum value. A strictly convex function has a unique global minimum[a].

find minimizer of

$$\min_{w} f(w)$$

Sums of convex functions are also convex. Therefore, MSE has only one global minimum value.

$$MSE(w) = \frac{1}{N}\sum_{n=1}^{N}(y_n - x_n^T w)^2$$

of a linear regression

[a] Read section 7.3.3 from Kevin Murphy's book for more details

Convexity is a desired *computational* property.

Can you prove that the MAE is convex? (as a function of the parameters $\mathbf{w}$)

Computational VS statistical trade-off

So which loss function is the best?

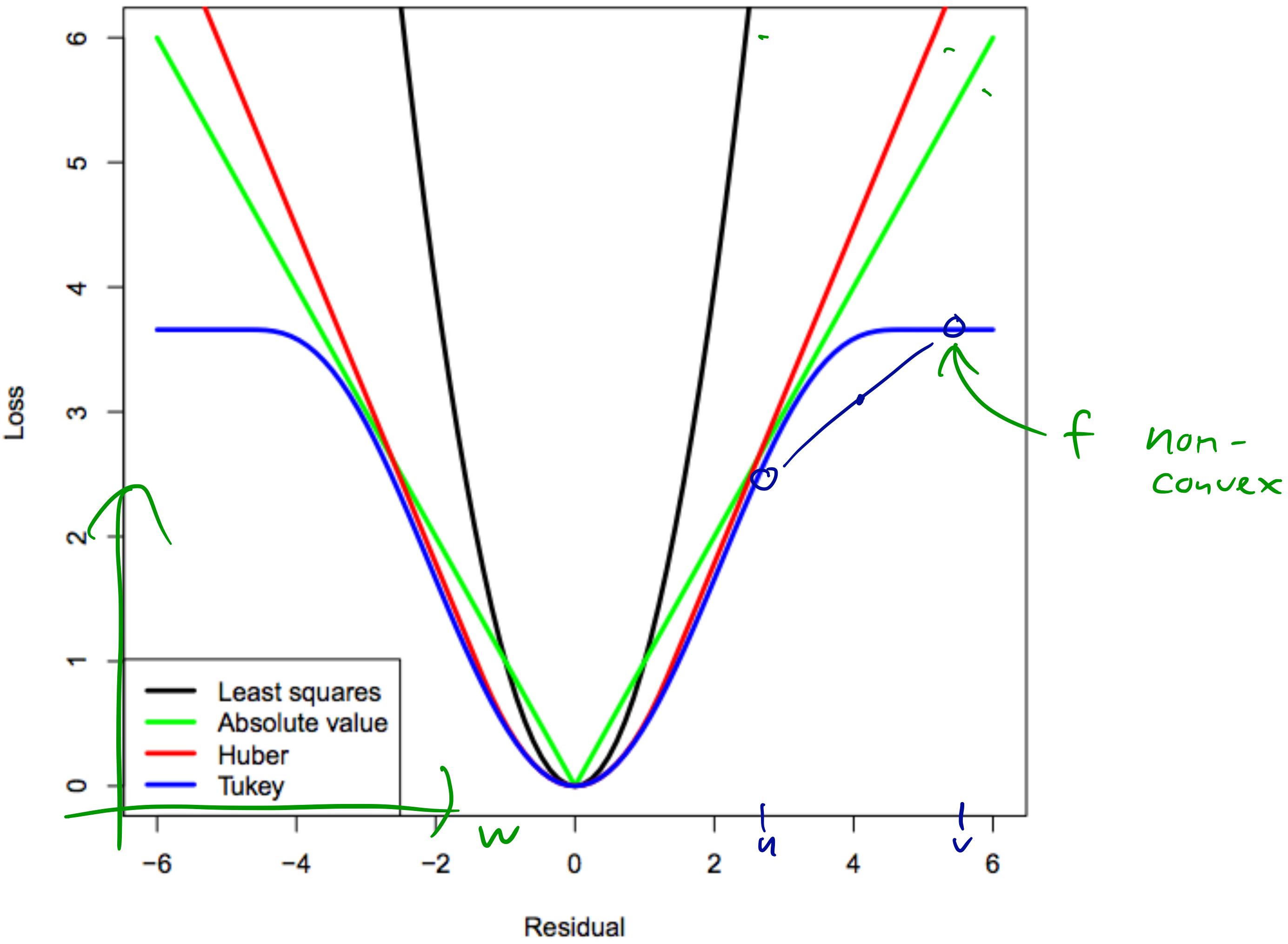


Figure taken from Patrick Breheny's slide.

If we want better statistical properties, then we have to give-up good computational properties.

Additional Reading

Other cost functions

Huber loss

$$Huber := \begin{cases} \frac{1}{2}e^2 & , \text{if } |e| \leq \delta \\ \delta|e| - \frac{1}{2}\delta^2 & , \text{if } |e| > \delta \end{cases} \tag{1}$$

Huber loss is convex, differentiable, and also robust to outliers. However, setting δ is not an easy task.

Tukey's bisquare loss (defined in terms of the gradient)

$$\frac{\partial \mathcal{L}}{\partial e} := \begin{cases} e\{1 - e^2/\delta^2\}^2 & , \text{if } |e| \leq \delta \\ 0 & , \text{if } |e| > \delta \end{cases} \tag{2}$$

Tukey's loss is non-convex, but robust to outliers.

Additional reading on convexity

- Read section 7.3.3 from Kevin Murphy's book for more details.
- Prove that the sum of two convex function is convex (Hint: use the definition).

Additional reading on outliers

- Read the Wikipedia page on "Robust statistics".
- Repeat the exercise with MAE.

A question for cost functions

Is there an automatic way to define cost functions?

Nasty cost functions: Visualization

See Andrej Karpathy Tumblr post for many cost functions gone "wrong" for neural network. `http://lossfunctions.tumblr.com/`.