Cost Functions

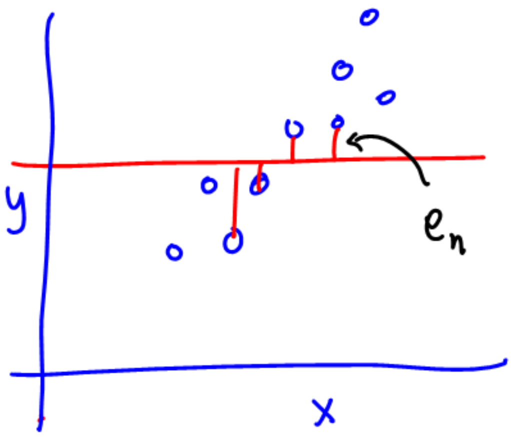
Mohammad Emtiyaz Khan
EPFL

Sep 17, 2015

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

Consider the following models.



1-parameter model: $y_n \approx \beta_0$
2-parameter model: $y_n \approx \beta_0 + \beta_1 x_{n1}$

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

What is a cost function?

Cost functions (or utilities or energy) are used to learn parameters that explain the data well. They define how costly our mistakes are.

Two desirable properties of cost functions

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

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

Statistical vs computational trade-off

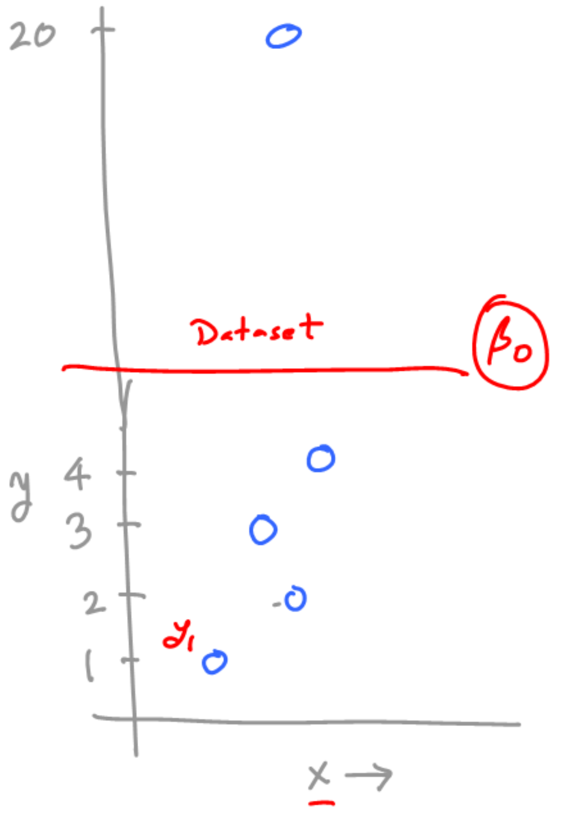
If we want better statistical properties, then we have to give good computational properties.

Mean square error (MSE)

MSE is one of the most popular cost function.

$$\color{blue}{MSE}(\boldsymbol{\beta}) := \sum_{n=1}^{N} [y_n - f(\mathbf{x}_n)]^2$$

Does it have both the properties?

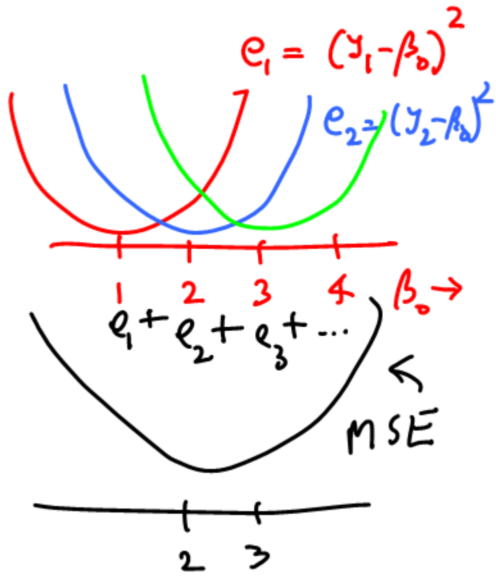


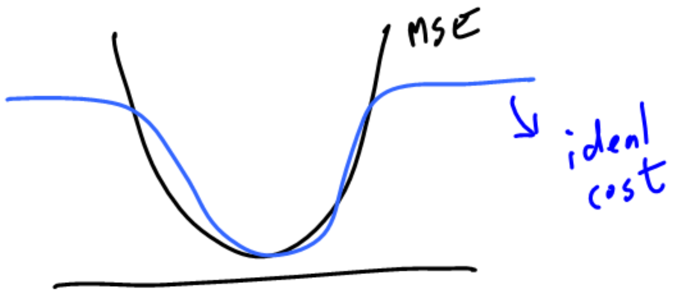
An exercise for MSE

Compute MSE for 1-param model:

$$\mathcal{L}(\beta_0) := \sum_{n=1}^{N} [y_n - \beta_0]^2 \qquad (1)$$

$\beta_0 \rightarrow$	1	2	3	4	5	6	7
$y_1 =$ 1	0	1	4	9	16	25	36
$y_2 =$ 2	1	0	1	4	9	16	25
$y_3 =$ 3	4	1	0	1	4	9	16
$y_4 =$ 4	9	4	1	0	1	4	9
MSE	14	6	6	14	30	54	86
$y_5 =$ 20							
MSE	375	330	295	270	255	250	255

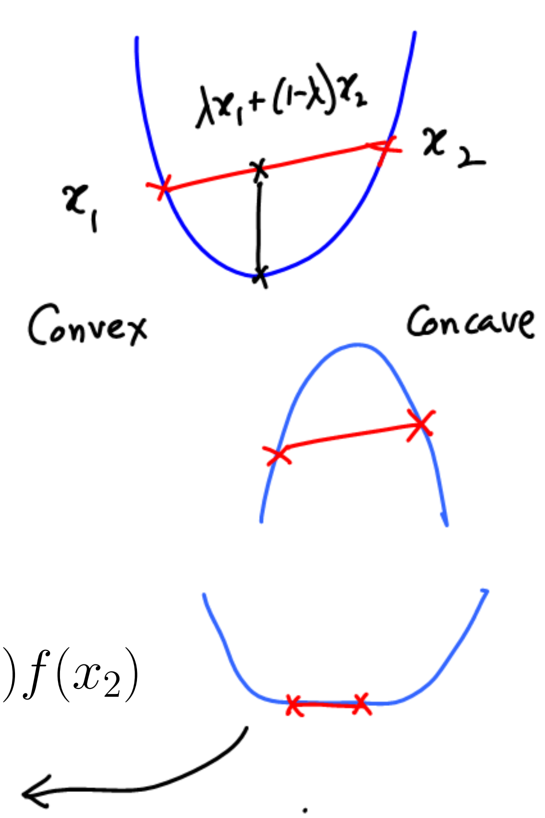




Some help: $19^2 = 361, 18^2 = 324, 17^2 = 289, 16^2 = 256, 15^2 = 225, 14^2 = 196, 13^2 = 169$.

Convexity

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



A function $f(x)$ with $x \in \mathcal{X}$ is convex, if for any $x_1, x_2 \in \mathcal{X}$ and for any $0 \leq \lambda \leq 1$, we have:

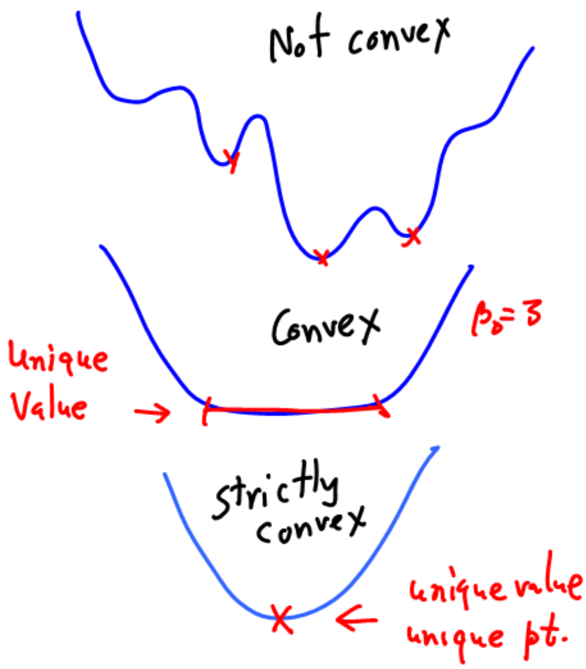
$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

A function is strictly convex if the inequality is strict.

Importance of convexity

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

No local minimum



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

Convexity is a desired *computational* property.

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

Outliers

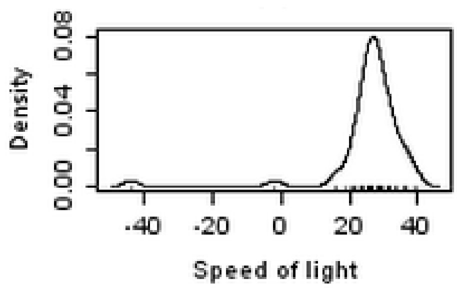
Outliers are data examples that are far away from most of the other examples. Unfortunately, they occur more often in reality than you would want them to!

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				

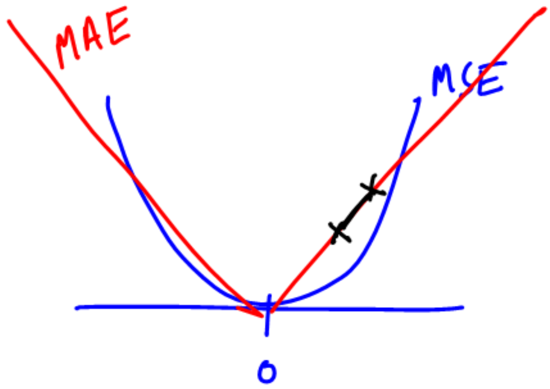
(a) Original speed of light data done by Simon Newcomb.



(b) Histogram showing outliers.

Handling outliers is a desired *statistical* property.

Mean Absolute Error (MAE)

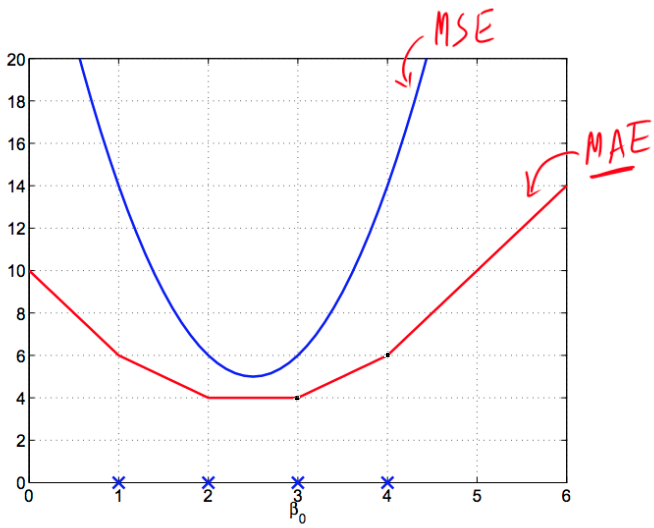


$$MAE := \sum_{n=1}^{N} |y_n - f(\mathbf{x}_{\cancel{i}\,n})| \qquad (2)$$

Repeat the exercise with MAE.

	1	2	3	4	5	6	7
1							
2							
3							
4							
MSE							
MSE							

What about convexity? Are there any issues? Can you draw MSE and MAE for the above example?



Computational Vs statistical trade-off

So which loss function is the best?

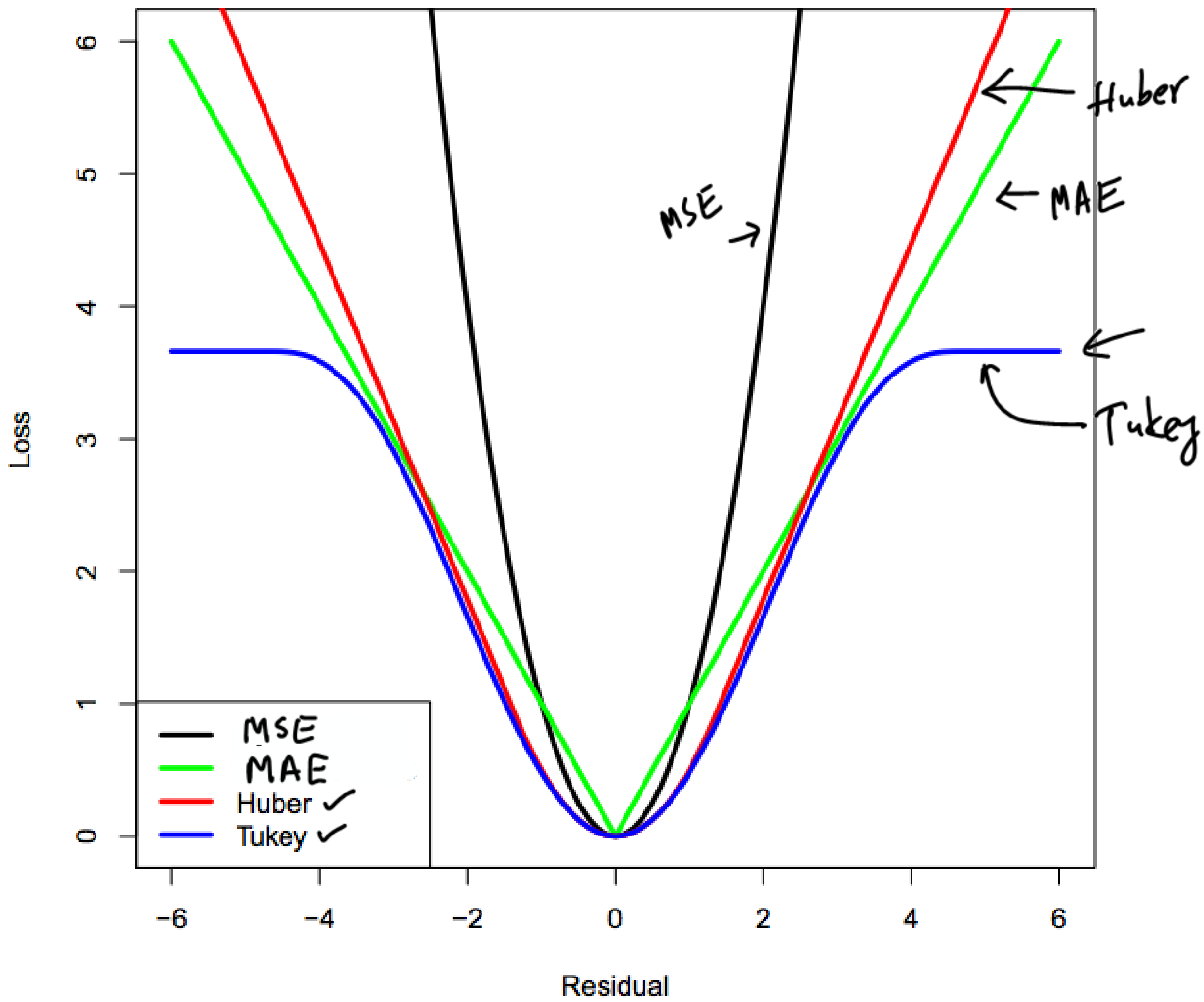


Figure is taken from Patrick Breheny's slide.

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

Additional Reading

Other cost functions

Huber loss

$$Huber := \begin{cases} \frac{1}{2}e^2 & \text{, if } |e| \leq \delta \\ \delta|e| - \frac{1}{2}\delta^2 & \text{, if } |e| > \delta \end{cases} \tag{3}$$

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 gradient)

$$\frac{\partial \mathcal{L}}{\partial e} := \begin{cases} e\{1 - e^2/\delta^2\}^2 & \text{, if } |e| \leq \delta \\ 0 & \text{, if } |e| > \delta \end{cases} \tag{4}$$

Tukey's loss is non-convex, non-differntiable, 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 for 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 loss functions?

Nasty cost functions: Visualization

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