Generalized Linear Models

Luca Fontana, Sanzio Monti

08.10.2018

- Introduction
- Logistic Regression
- Multiclass Logistic Regression
- Poisson GLM
- Cox Proportional Hazard Models
- Support Vector Machines

In **linear regression:** continuous response $Y \in \mathbb{R}$ with $\hat{y} = f(x) = \beta_0 + \beta^\top x \in \mathbb{R}^n$, Gaussian error.

Idea: Apply linear regression on "smartly" transformed variable of interest.
In this way we can work with, for example, binary or counts responses.

The **link function** g is a strictly monotonic transformation of the conditional mean of Y given X:

$$\mu(x) = \mathbb{E}[Y \mid X = x]$$
$$g[\mu(x)] = \beta_0 + \beta^\top x$$

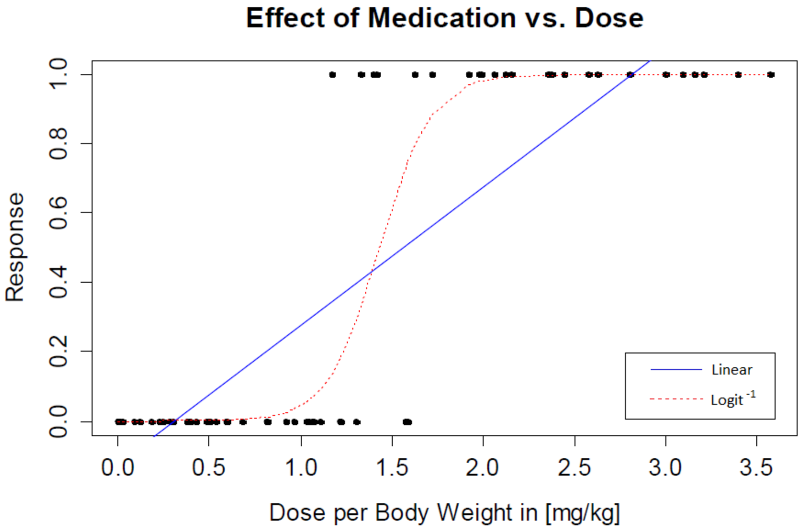
$Y \in \{0, 1\}$,

$$\begin{aligned}\mu(x) &= \mathbb{E}[Y|X = x] \\ &= 1 \cdot \mathbb{P}(Y = 1 \mid X = x) + 0 \cdot \mathbb{P}(Y = 0 \mid X = x) \\ &= \mathbb{P}(Y = 1 \mid X = x)\end{aligned}$$

Take $g(\mu) = \mathrm{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$

$$\beta_0 + \beta^\top x = g[\mu(x)] = \log\left[\frac{\mathbb{P}(Y = 1 \mid X = x)}{\mathbb{P}(Y = 0 \mid X = x)}\right]$$

$$\mu(x) = \mathbb{P}(Y = 1 \mid X = x) = \frac{e^{\beta_0 + \beta^\top x}}{1 + e^{\beta_0 + \beta^\top x}}$$



Source: Marcel Dettling, lecture notes of Applied Statistical Regression lecture, Fall Semester 2017, page 39.

$Y \in \mathbb{N} \subset [0, +\infty[$, Assume that $(Y \mid X = x) \sim Poi(\lambda(x))$.

$$\mu(x) = \mathbb{E}[Y|X = x] = \lambda(x)$$

Take $g(\mu) = \log(\mu)$

$$\beta_0 + \beta^\top x = g[\mu(x)] = \log(\mu(x)) = \log(\lambda(x))$$

$$\lambda(x) = e^{\beta_0 + \beta^\top x}$$

Problem in chapter: Minimize negative log-likelihood with a penalty.

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ -\frac{1}{N}\mathcal{L}(\beta_0, \beta; \boldsymbol{y}, \boldsymbol{X}) + \lambda \left\| \beta \right\| \right\}$$

Were the type of norm is specified in the problem.

We now show that linear regression is an example of GLM. Assume $(Y \mid X = x) \sim \mathcal{N}(\mu(x), \sigma^2)$. Then we have:

$$\tilde{\mathcal{L}}(\beta_0, \beta; \boldsymbol{y}, \boldsymbol{X}) \propto \prod_{i=1}^{N} e^{-\frac{(y_i - \beta_0 - \beta x_i)^2}{2\sigma^2}}$$

$$\mathcal{L}(\beta_0, \beta; \boldsymbol{y}, \boldsymbol{X}) = -\sum_{i=1}^{N} \frac{(y_i - \beta_0 - \beta x_i)^2}{2\sigma^2} + c = -\frac{\|\boldsymbol{y} - \beta_0 - \beta \boldsymbol{X}\|_2^2}{2\sigma^2} + c$$

Binary $Y \in \{0, 1\}$: assume Bernoulli distributed with parameter $\mu(x) = P(Y = 1 \mid X = x)$; Corresponding likelihood:

$$\tilde{\mathcal{L}}(\beta_0, \beta; \boldsymbol{x}) = \prod_{i=1}^{N} \underbrace{\mathbb{P}(Y = 1 \mid X = x_i)}_{p(x_i)}^{y_i} \mathbb{P}(Y = 0 \mid X = x_i)^{1-y_i}$$

$$\mathcal{L}(\beta_0, \beta; \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{N} y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))$$

$$\underset{\beta_0, \beta}{\text{minimize}} \left\{ -\frac{1}{N} \mathcal{L}(\beta_0, \beta; \boldsymbol{y}, \boldsymbol{X}) + \lambda \|\beta\| \right\}$$

Binary $Y \in \{0,1\}$: assume Bernoulli distributed with parameter $\mu(x) = P(Y=1 \mid X=x)$; Corresponding likelihood:

$$\tilde{\mathcal{L}}(\beta_0, \beta; \boldsymbol{x}) = \prod_{i=1}^{N} \underbrace{\mathbb{P}(Y=1 \mid X=x_i)}_{p(x_i)}^{y_i} \mathbb{P}(Y=0 \mid X=x_i)^{1-y_i}$$

$$\mathcal{L}(\beta_0, \beta; \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{N} y_i \log(p(x_i)) + (1-y_i)\log(1-p(x_i))$$

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ -\frac{1}{N} \sum_{i=1}^{N} y_i \log(p(x_i)) + (1-y_i)\log(1-p(x_i)) + \lambda \|\beta\| \right\}$$

Using the Ansatz $p(x) = \mathbb{P}(Y = 1 \mid X = x) = \frac{e^{\beta_0 + \beta^\top x}}{1 + e^{\beta_0 + \beta^\top x}}$ we get:

$$\underbrace{-\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + \beta^\top x_i) - \log\left(1 + e^{\beta_0 + \beta^\top x_i}\right) + \lambda \|\beta\|}_{\text{minimize}}$$

Sometimes we take $Y \in \{-1, 1\}$, which simplifies the minimization problem to:

$$\frac{1}{N}\sum_{i=1}^{N} \log\left(1 + e^{-y_i(\beta_0 + \beta^\top x_i)}\right) + \lambda \|\beta\|$$

or more generally:

$$\frac{1}{N}\sum_{i=1}^{N} \log\left(1 + e^{-y_i f(x_i, \beta_0, \beta)}\right) + \lambda \|\beta\|$$

Problem: We have $N = 11'314$ documents that we want to classify into two different Groups ($Y \in \{-1, +1\}$). The features are defined as the set of **trigrams** [with some restrictions]. Trigrams are a sequence of three consecutive characters (for example AAA, azA,...). Each document contains an average of 425 nonzero features.

$$p = 777'811 \cong 92^3$$

We want to perform ℓ_1-regularized logistic regression.

Example: Document classification

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Example: Document classification

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

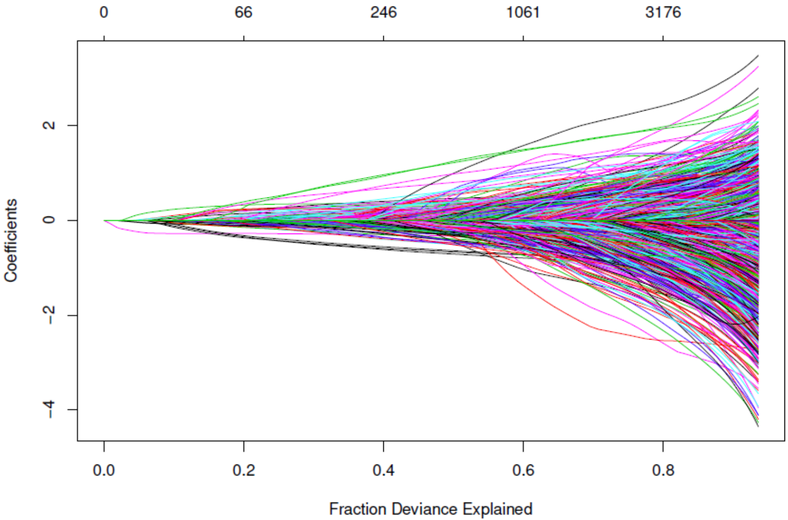
The **fraction deviance explained** (D^2_λ) is then defined by:

$$D^2_\lambda = \frac{\mathrm{Dev}_{\mathrm{null}} - \mathrm{Dev}_\lambda}{\mathrm{Dev}_{\mathrm{null}}}$$

$$R^2 = \frac{\mathrm{SS}_{\mathrm{tot}} - \mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$

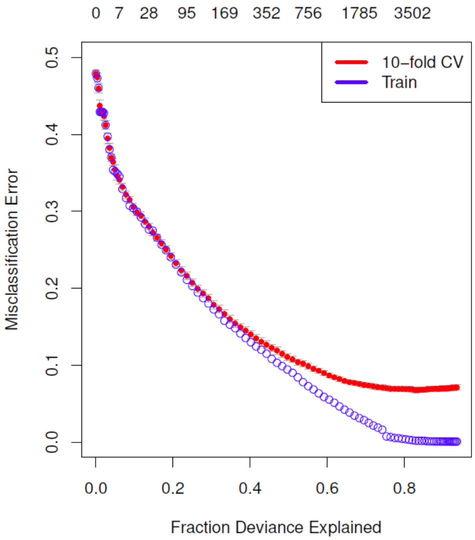
Deviance: (Dev_λ) Minus twice the difference in log-likelihood of a saturated model with a model fit with parameter λ.
Goodness to fit statistic that generalizes the residual sum of square for cases where model fitting is achieved by MLE.

Example: Document classification



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the Lasso and generalizations. CRC Press, 2015, page 33.

Example: Document classification



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the Lasso and generalizations. CRC Press, 2015, page 34.

In order to calculate the coefficients, the package `glmnet` approximates the log-likelihood:

$$\frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{-y_i(\beta_0+\beta^\top x_i)}\right)$$

in each step by a quadratic function, in order to reuse the method created to solve the "usual" ridge and lasso regression. More on that will be covered in future presentations.

Setting: $Y \in \{1, \ldots, K\}$ for $K > 2$ classes. 2 ways for reduction to binary classification in general:

- OvO [**One versus One**]: all $\binom{K}{2}$ pairs of classes samples are used to fit $\binom{K}{2}$ binary classifiers, then the predicted class is the one which is predicted the most.
- OvA [**One versus All**]: treat all other classes as a single *negative* class.

Drawbacks:

- OvO: computationally exhaustive and cases where same amount of votes for more classes.
- OvA: imbalance amounts *positive* and *negative* observations.

Multiclass Logistic Regression approach:

$$\mathbb{P}(Y = k \mid X = x; \beta_0, \beta) = \frac{e^{\beta_{0k} + \beta_k^\top x}}{\sum_{l=1}^{K} e^{\beta_{0l} + \beta_l^\top x}}$$

Interesting property: invariance of this conditional probability under addition of $\gamma_0 + \gamma^\top x$ in all the exponents.

$$\log \mathbb{P}(Y = k \mid X = x; \beta_0, \beta) = \beta_{0k} + \beta_k^\top x - \log \left[\sum_{l=1}^{K} e^{\beta_{0l} + \beta_l^\top x} \right]$$

The log-likelihood is given by

$$\mathcal{L}(\beta_0, \beta; \boldsymbol{y}, \boldsymbol{X}) = \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{P}(Y = y_i \mid X = x_i; \beta_0, \beta)$$

Define **Indicator response** $\boldsymbol{R} = (r_{ik}) = \left(\mathbb{I}_{\{y_i = k\}}\right) \in \mathbb{R}^{N \times K}$.

Then we can rewrite log-likelihood as

$$\frac{1}{N} \sum_{i=1}^{N} {\color{red} w_i} \left[\sum_{k=1}^{K} r_{ik}(\beta_{0k} + \beta_k^\top x_i) - \log \left\{ \sum_{k=1}^{K} e^{\beta_{0k} + \beta_k^\top x_i} \right\} \right]$$

with resulting ℓ_1-penalized negative log-likelihood

$$-\frac{1}{N} \sum_{i=1}^{N} w_i \left[\sum_{k=1}^{K} r_{ik}(\beta_{0k} + \beta_k^\top x_i) - \log \left\{ \sum_{k=1}^{K} e^{\beta_{0k} + \beta_k^\top x_i} \right\} \right] + \lambda \sum_{k=1}^{K} \|\beta_k\|_1$$

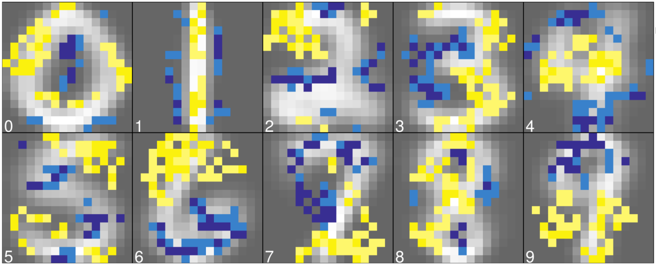
Recall invariance under addition of linear term in exponents: $\{\tilde{\beta}_{kj}\}_{k=1}^{K}$ and $\{\tilde{\beta}_{kj} + c_j\}_{k=1}^{K}$ produce same probabilities, $c_j \in \mathbb{R}$.

Use penalty $\sum_{k=1}^{K} \|\tilde{\beta}_k\|_1 = \sum_{k=1}^{K} \sum_{j=1}^{p} |\tilde{\beta}_{kj}| = \sum_{j=1}^{p} \sum_{k=1}^{K} |\tilde{\beta}_{kj}|$ to choose $\{c_j\}_{j=1}^{p}$.

For all candidates $\{\tilde{\beta}_{kj}\}_{k=1}^{K}$, optimal $c_j \in \mathbb{R}$ satisfies

$$c_j = \arg\min_{c \in \mathbb{R}} \sum_{k=1}^{K} |\tilde{\beta}_{kj} - c|, \quad j \in \{1, \ldots, p\}$$

Solution: $c_j = \text{median}\{\tilde{\beta}_{1j}, \ldots, \tilde{\beta}_{Kj}\}$, for $j \in \{1, \ldots, p\}$.



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the Lasso and generalizations. CRC Press, 2015, page 38.

Problem: We have $N = 7'921$ gray-scale images of $p = 256$ pixels representing **handwritten digits** from 0 to 9 ($Y \in \{0, \dots, 9\}$).
Each one of the p features represents the intensity in a $[0, 1]$-scale of the corresponding pixel (0 *black*, 1 *white*).
We can fit a $10-$classes lasso multinomial model.

We can introduce sparsity via grouped-lasso penalty: consider the vector of class-coefficients for feature j

$$\boldsymbol{\beta}_j = (\underbrace{\beta_{0j}}_{\text{digit } 0}, \ldots, \underbrace{\beta_{9j}}_{\text{digit } 9}) \quad j \in \{1, \ldots, p\}$$

Replace standard multinomial criterion with grouped-lasso one:

$$-\frac{1}{N}\sum_{i=1}^{N} \log \mathbb{P}(Y = y_i \mid X = x_i; \{\boldsymbol{\beta}_j\}_{j=1}^{p}) + \lambda \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_2$$

Consequence: all coefficients of a particular feature are in or out of the model, i.e. the feature is in or out of the model.

Setting: Y non-negative and represents a count.

Approach: Poisson likelihood and log-linear model for mean

$$\log \mu(x) = \beta_0 + \beta^\top x$$

with resulting ℓ_1-penalized negative log-likelihood

$$-\frac{1}{N} \sum_{i=1}^{N} \left\{ y_i(\beta_0 + \beta^\top x_i) - e^{\beta_0 + \beta^\top x_i} \right\} + \lambda \left\| \beta \right\|_1$$

we aim to minimize.

Typical situation: model rates [e.g. death rate] via Poisson model. If observation windows have different lengths T_i, then

$$\mathbb{E}[y_i \mid X_i = x_i] = T_i \mu(x_i)$$

where $\mu(x_i)$ rate per unit time interval.
Example: 6 months vs yearly visit to doctor has $T = 1/2$.

New model form:

$$\log \mathbb{E}[Y \mid X = x, T] = \underbrace{\log T}_{\text{"offset"}} + \beta_0 + \beta^\top x$$

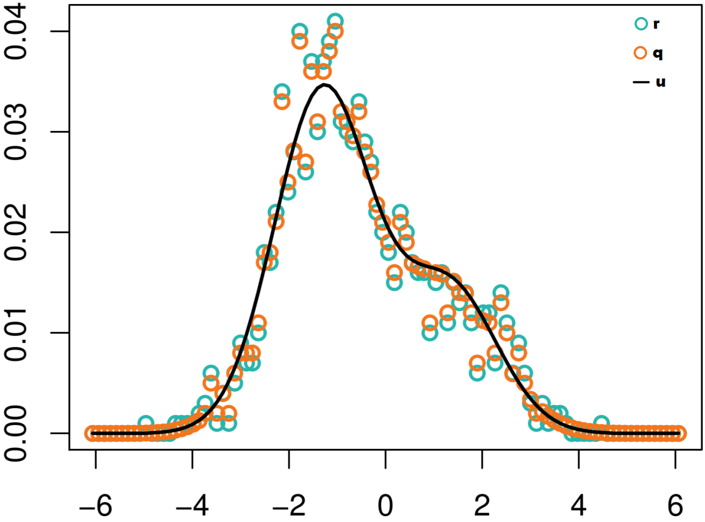
Problem: N count variables $\{y_k\}_{k=1}^N$ coming from a N-cell multinomial distribution.

$\boldsymbol{r} = \{r_k\}_{k=1}^N = \{y_k / \sum_{k=1}^N y_k\}_{k=1}^N$ vector of proportions.

Issue: $\boldsymbol{r}$ could be sparse. Want to regularize it toward a more stable distribution $\boldsymbol{u} = \{u_k\}_{k=1}^N$.

$$\underset{\boldsymbol{q} \in \mathbb{R}^N,\, q_k \geq 0}{\text{minimize}} \quad \underbrace{\sum_{k=1}^N q_k \log\left(\frac{q_k}{u_k}\right)}_{\text{Kullback-Leibler divergence}} \quad \text{such that } \|\boldsymbol{q} - \boldsymbol{r}\|_\infty \leq \delta, \sum_{k=1}^N q_k = 1$$

We want a distribution $\boldsymbol{q}$ which is approximately equal to our observed proportions but at the same time as close as possible to a nominal distribution $\boldsymbol{u}$.



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the Lasso and generalizations. CRC Press, 2015, page 41.

Why this problems falls in Poisson model framework?
The previous minimization problem

$$\underset{\boldsymbol{q}\in\mathbb{R}^N,\, q_k\geq 0}{\text{minimize}} \sum_{k=1}^{N} q_k \log\left(\frac{q_k}{u_k}\right) \text{ such that } \|\boldsymbol{q}-\boldsymbol{r}\|_\infty \leq \delta,\ \sum_{k=1}^{N} q_k = 1$$

has Lagrange dual

$$\underset{\beta_0,\boldsymbol{\alpha}}{\text{maximize}} \left\{ \sum_{k=1}^{N} r_k \left[\log u_k + \beta_0 + \alpha_k - u_k e^{\beta_0+\alpha_k}\right] - \delta \|\boldsymbol{\alpha}\|_1 \right\}$$

This is equivalent to fitting a Poisson model with offset $\log u_k$, individual parameter α_k and design matrix $X = \mathbb{I}_{N\times N}$.

Setting: Medical studies interested in time to death T of sick patients, usually characterized by the survivor function $S(t) := \mathbb{P}(T > t)$, the probability of surviving beyond a certain time t.

Some patients drop out the study or die because of unrelated causes: we call this situation a *censoring* time C.

$Y := \min(C, T)$ is the observed variable, together with an indicator $\delta := \mathbb{I}_{\{Y=T\}}$ of whether the patient died *correctly* (because of the studied illness).

Hazard function: *Instantaneous probability of death at time t, given survival up till t.*

$$h(t) = \lim_{\delta \to 0} \frac{\mathbb{P}(Y \in \{t, t+\delta\} \mid Y \geq t)}{\delta} = \frac{f(t)}{S(t)}$$

where $f(t)$ density of T.

Cox's model treats special cases of hazard functions:

$$h(t; x) = h_0(t) e^{\beta^\top x}$$

where x represents e.g. gene expressions and $h_0(t)$ is **baseline hazard**: hazard for one individual with $x = 0$.

$$h(t;x) = h_0(t)e^{\beta^\top x}$$

Denote by $R_i := \{j \mid y_j \geq y_i\}$ the **risk set** of subject i (individuals which are still in the study when subject i dies), then the partial likelihood of subject i is given by

$$\mathbb{P}(Y_i = y_i) = \frac{h(y_i;x_i)}{\sum_{j \in R_i} h(y_j;x_j)} = \frac{e^{\beta^\top x_i}}{\sum_{j \in R_i} e^{\beta^\top x_j}}$$

Note that baseline hazard h_0 has no effect here.

The log-Likelihood is

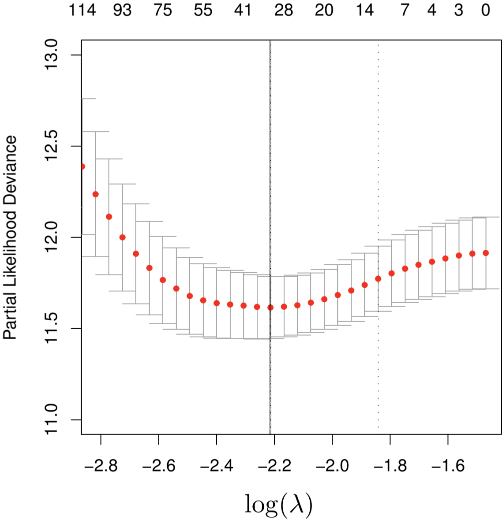
$$\mathcal{L}(\beta; \boldsymbol{x}, \boldsymbol{\delta}) = \sum_{\underbrace{i \,:\, \delta_i = 1}_{\text{died "correctly"}}} \log \left[\frac{e^{\beta^\top x_i}}{\sum_{j \in R_i} e^{\beta^\top x_j}} \right]$$

with corresponding ℓ_1-penalized CPH problem:

$$\underset{\beta}{\text{minimize}} \left\{ - \sum_{i:\delta_i=1} \log \left[\frac{e^{\beta^\top x_i}}{\sum_{j \in R_i} e^{\beta^\top x_j}} \right] + \lambda \left\| \beta \right\|_1 \right\}$$

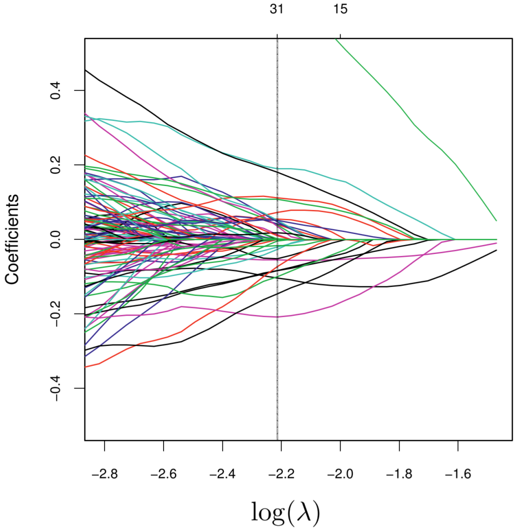
Problem: We want to estimate the survivor function S for $N = 240$ Lymphoma patients with $p = 7399$ variables measuring gene expressions. 102 of these samples are *right censored*, i.e. $Y = \min(T, C) = C$.

We select $\lambda_{\min}$ via CV using the deviance.



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the Lasso and generalizations. CRC Press, 2015, page 45.

We use the ℓ_1-penalized CPH problem to find $\hat{\beta}(\lambda_{\min})$



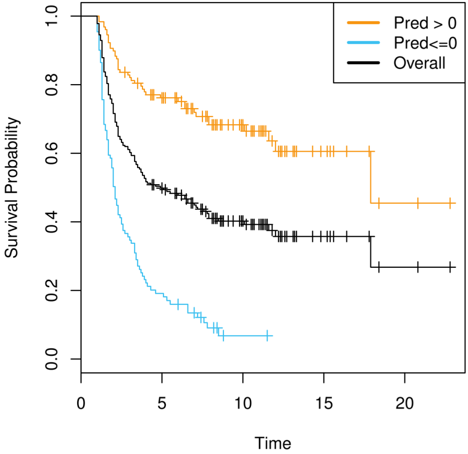
Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the Lasso and generalizations. CRC Press, 2015, page 44.

We use the **Kaplan-Meier estimator** of survivor function $S(t)$: let $\hat{\eta}(x) := \hat{\beta}(\lambda_{\min})^\top x$, then

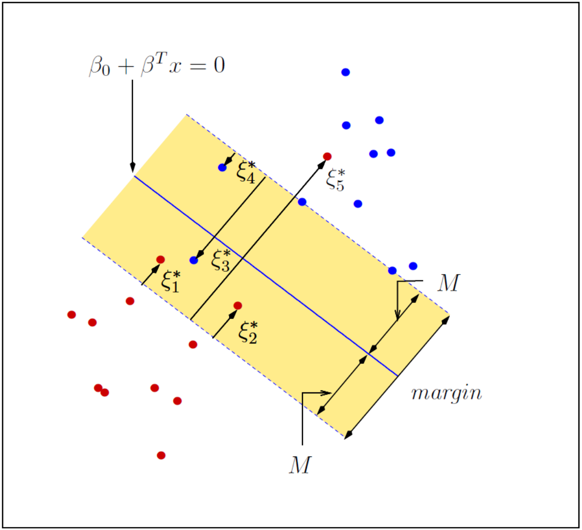
$$\widehat{S}(t) = \prod_{i:y_i \le t} \left(1 - \frac{e^{\hat{\eta}(x_i)}}{\sum_{j \in R_i} e^{\hat{\eta}(x_j)}}\right)$$

is an estimate of $S(t)$. We use these in the following plot.

We create two groups $\{i : \hat{\eta}(x_i) > 0\}$ and $\{i : \hat{\eta}(x_i) \leq 0\}$, then we compute the estimate using them and the overall set, resulting in three curves.



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the Lasso and generalizations. CRC Press, 2015, page 42.

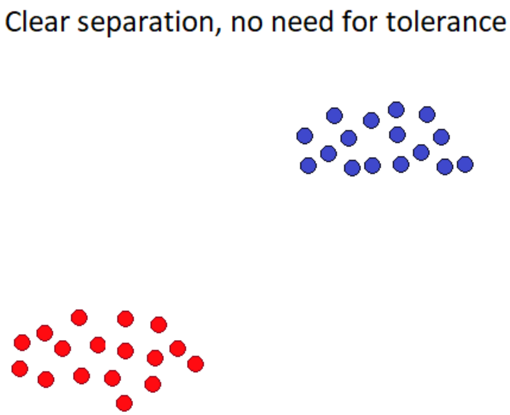


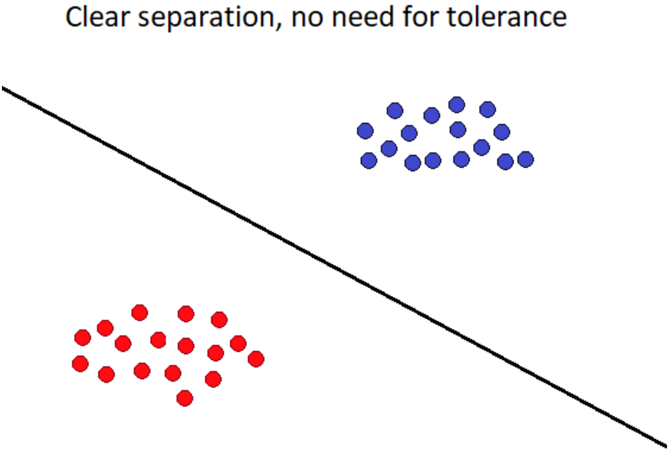
Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the Lasso and generalizations. CRC Press, 2015, page 47.

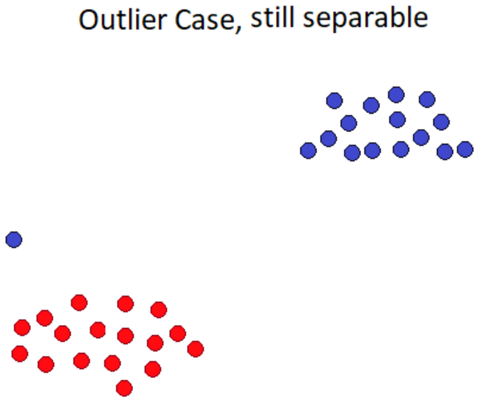
$$\underset{\beta_0,\beta,\{\xi_i\}_1^N}{\text{maximize}}\, M \text{ subject to } y_i\underbrace{(\beta_0 + \beta^\top x_i)}_{f(x_i,\beta_0,\beta)} \geq M(1-\xi_i)\forall i$$

and

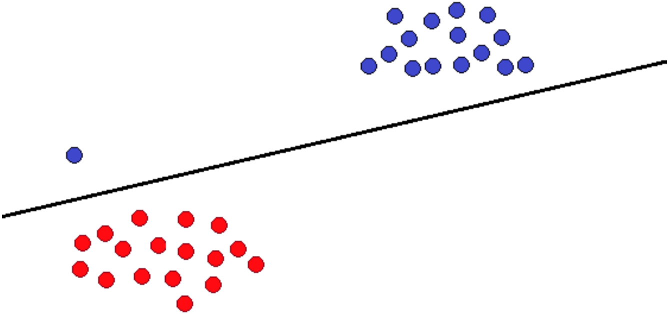
$$\xi_i \geq 0\ \forall i, \sum_{i=1}^{N} \xi_i \leq C, \|\beta\|_2 = 1$$


Clear separation, no need for tolerance

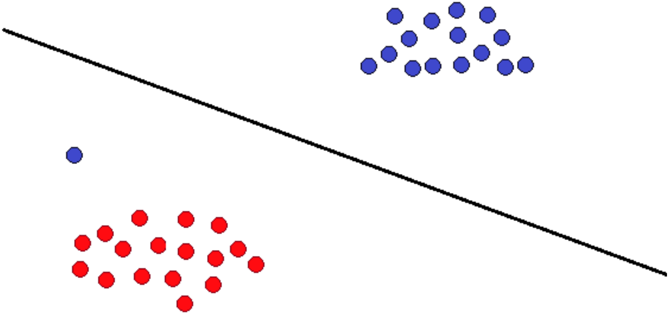

Clear separation, no need for tolerance


Outlier Case, still separable

Outlier Case, C small



Outlier Case, C Big



The constraint can be seen as:

$$\begin{cases} f(x_i, \beta_0, \beta) \geq +1, & \text{when } y_i = +1 \\ f(x_i, \beta_0, \beta) \leq -1, & \text{when } y_i = -1 \end{cases}$$

This can be rewritten as $y_i f(x_i, \beta_0, \beta) \geq 1$ or

$$0 \geq 1 - y_i f(x_i, \beta_0, \beta)$$

By writing the Lagrangian equivalent of the original minimization problem (SVM), we get:

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} [1 - y_i f(x; \beta_0, \beta)]_+ + \lambda \left\| \beta \right\|_2^2 \right\}$$

Decreasing λ corresponds to decreasing C.

We now want to compare ridge penalized logistic regression:

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \underbrace{\log(1 + e^{-y_i f(x_i,\beta_0,\beta)})}_{\text{logistic loss}} + \lambda \|\beta\|_2^2 \right\}$$

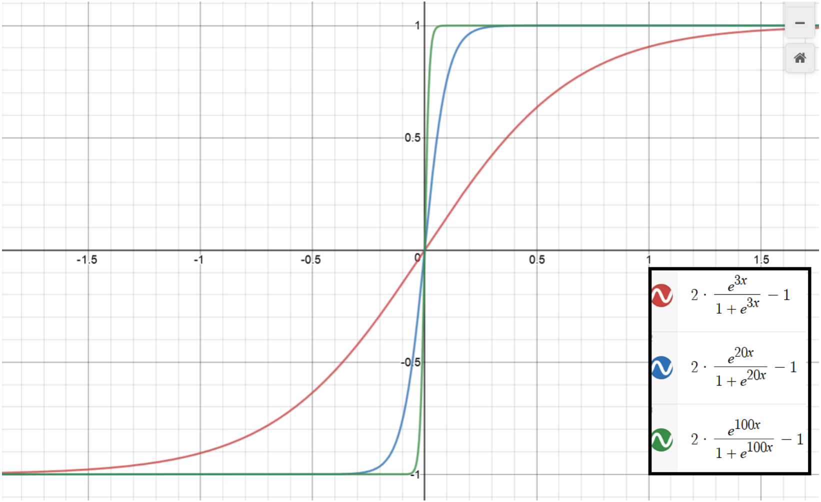
with the SVM problem:

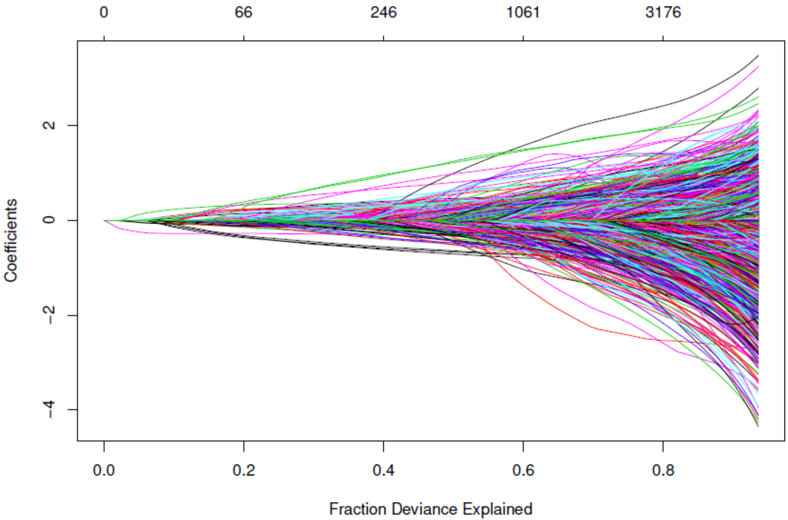
$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \underbrace{[1 - y_i f(x; \beta_0, \beta)]_+}_{\text{hinge loss}} + \lambda \|\beta\|_2^2 \right\}$$

Data is **separable**: there exists a hyperplane that separates the two cases. In this cases logistic regression has a problem:

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{e^{\beta_0 + \beta^\top x}}{1 + e^{\beta_0 + \beta^\top x}}$$

Problem: When $p >> N$, the points are almost always separable.


1
0.5
-0.5
-1
-1.5
-1
-0.5
0
0.5
1
1.5
$2 \cdot \frac{e^{3x}}{1+e^{3x}} - 1$
$2 \cdot \frac{e^{20x}}{1+e^{20x}} - 1$
$2 \cdot \frac{e^{100x}}{1+e^{100x}} - 1$



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the Lasso and generalizations. CRC Press, 2015, page 33.

From logistic regression we can construct a **linear classifier** in order to compare it with SVM:
If $\mathbb{P}(Y = -1 \mid X = x_{\text{observed}}) > 0.5$ then $y_{\text{predicted}} = -1$ and vice versa.
This is linear, because :

$$\mathbb{P}(Y = -1 \mid X = x) = \frac{1}{1 + e^{\beta_0 + \beta^\top x}} = \frac{1}{2} \iff e^{\beta_0 + \beta^\top x} = 1$$

$$\beta_0 + \beta^\top x = \log(1) = 0$$

Consider the Boundary $B = \{x \in \mathbb{R}^p | f(x) = 0\}$, where

$$f(x) = \beta_0 + \beta^\top x$$

Then the distance between the boundary and the point x_0 is

$$\operatorname{dist}(x_0, B) = \inf_{z \in B} \|z - x_0\|_2 = \frac{|\ f(x_0)\ |}{\|\beta\|_2}$$

So we find that the optimal separating plane $f^*(x) = 0$ has margin

$$M_2^* = \max_{\beta_0, \beta} \left\{ \min_{i \in \{1, \ldots, n\}} \frac{y_i f(x_i, \beta_0, \beta)}{\|\beta\|_2} \right\}$$

Consider the problem

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-y_i f(x_i,\beta_0,\beta)}) + \lambda \left\| \beta \right\|_2^2 \right\}$$

Let $(\tilde{\beta}_0(\lambda), \tilde{\beta}(\lambda))$ be the solution, then

$$M_2^* = \lim_{\lambda \to 0} \left\{ \min_{i \in \{1,\ldots,N\}} \frac{y_i f(x_i, \tilde{\beta}_0(\lambda), \tilde{\beta}(\lambda))}{\left\| \tilde{\beta}(\lambda) \right\|_2} \right\}$$

So for $\lambda \to 0$ we have that the ℓ_2-regularized logistic regression corresponds to the SVM solution.

In particular, if $(\breve{\beta}_0, \breve{\beta})$ solve the SVM problem for $C = 0$, then we have that:

$$\lim_{\lambda \to 0} \frac{\tilde{\beta}(\lambda)}{\left\| \tilde{\beta}(\lambda) \right\|_2} = \breve{\beta}$$

Note that the division by the ℓ_2 norm of $\tilde{\beta}(\lambda)$ makes sure that the solution on the SVM problem does not blow up.

To summarize:

- As $\lambda \to 0$, logistic regression and SVM solutions coincide.
- SVM leads to a more stable numerical method for computing the solution in this region.
- Logistic regression is more useful in the sparser part of the solution path.

Source: https://blog.phenoswitchbioscience.com/hubfs/question-mark-icon-pT5eMekqc.png?t=1530805015566

- Marcel Dettling, *Applied Statistical Regression*, Fall Semester 2017.
- Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the Lasso and generalizations. CRC Press, 2015, Chapter 3.