Generalised Linear Models

Dr. Robert Loftin

COM6012 Scalable Machine Learning
26.02.2024

Thanks to Dr. Shuo Zhou for these slides

Contents

Introduction

- Linear regression, $y \rightarrow$ continuous.

$$\mathbb{E}[y|\mathbf{x}] = \mathbf{w}^\top \mathbf{x},$$

- Logistic regression, $y \rightarrow$ binary ($y \in \{0, 1\}$).

$$\mathbb{E}[y|\mathbf{x}] = \frac{\exp^{\mathbf{w}^\top \mathbf{x}}}{1 + e^{\mathbf{w}^\top \mathbf{x}}},$$

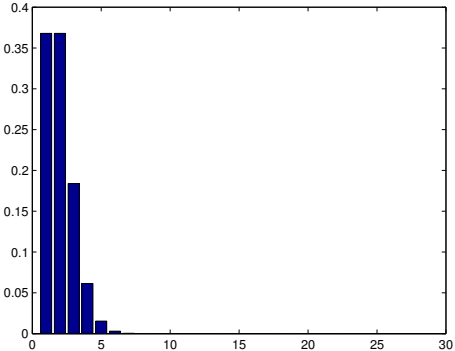
- $y \rightarrow$ non-negative integers, e.g. number of traffic incidents at an intersection in a year?

Poisson Distribution

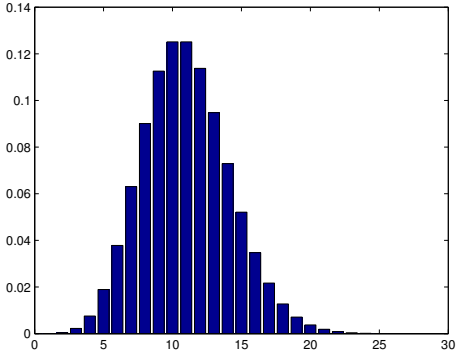
Distribution of non-negative integers y, with probability mass function

$$\text{Poi}(y|\mu) = \exp(-\mu)\frac{\mu^y}{y!}.$$

Here $\mu > 0$ is called the *rate parameter*, the average number of *events* in a given period.



$\mu = 1$



$\mu = 10$

Poisson Regression

- Assume that:

$$p(y_n|\mathbf{x}_n, \mathbf{w}) = \text{Poi}(y_n|\mu(\mathbf{x}_n)) = \exp(-\mu_n)\frac{\mu_n^y}{y!},$$

  where $\mu_n = \exp(\mathbf{w}^\top \mathbf{x}_n)$

- The log probability mass function is then:

$$\begin{aligned}\log p(y_n|\mathbf{x}_n, \mathbf{w}) &= y_n \log \mu_n - \mu_n - \log(y_n!) \\ &= y_n \mathbf{w}^\top \mathbf{x}_n - \exp(\mathbf{w}^\top \mathbf{x}_n) - \log(y_n!)\end{aligned}$$

- and the log-likelihood of a dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$:

$$p(y_1, y_2, \ldots, y_N|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N; \mathbf{w}) = \prod_{i=1}^{N} \exp(-\exp(\mathbf{w}^\top \mathbf{x}_i))\frac{\exp(y_i \mathbf{w}^\top \mathbf{x}_i)}{y_i!}$$

Generalised Form?

- Linear regression:

$$\mathbb{E}[y|\mathbf{x}] = \mathbf{w}^\top \mathbf{x}$$

- Logistic regression:

$$\mathbb{E}[y|\mathbf{x}] = \frac{\exp^{\mathbf{w}^\top \mathbf{x}}}{1 + e^{\mathbf{w}^\top \mathbf{x}}}$$

- Poisson regression

$$\mathbb{E}[y|\mathbf{x}] = e^{\mathbf{w}^\top \mathbf{x}}$$

- General form?

$$g(\mathbb{E}[y|\mathbf{x}]) = \mathbf{w}^\top \mathbf{x} \quad \text{or} \quad \mathbb{E}[y|\mathbf{x}] = g^{-1}(\mathbf{w}^\top \mathbf{x})$$

Contents

Exponential Families

- A PDF or PMF $p(\mathbf{y}|\boldsymbol{\eta})$ is a member of the **exponential family** if it can be written as

$$p(\mathbf{y}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{y}) \exp\left[\boldsymbol{\eta}^\top \mathcal{T}(\mathbf{y})\right],$$

  where $Z(\boldsymbol{\eta}) = \int h(\mathbf{y}) \exp\left[\boldsymbol{\eta}^\top \mathcal{T}(\mathbf{y})\right] d\mathbf{y}$.

- $\boldsymbol{\eta} \in \mathbb{R}^d$ are known as the **natural parameters** or **canonical parameters** of the distribution.
- $\mathcal{T}(\mathbf{y}) \in \mathbb{R}^d$ is a **sufficient statistic** of y with respect to $\boldsymbol{\eta}$.
- $Z(\boldsymbol{\eta})$ is known as the **partition function**.
- $h(\mathbf{y})$ is a scaling constant (independent of $\boldsymbol{\eta}$).

Exponential Families (II)

- Distributions in the exponential family can also be expressed as

$$p(\mathbf{y}|\boldsymbol{\eta}) = h(\mathbf{y}) \exp\left[\boldsymbol{\eta}^\top \mathcal{T}(\mathbf{y}) - A(\boldsymbol{\eta})\right],$$

  where

$$A(\boldsymbol{\eta}) = \log Z(\boldsymbol{\eta}).$$

- $A(\boldsymbol{\eta})$ is called the **log partition function** or **cumulant function** (we'll see why later).

- When $\mathcal{T}(\mathbf{y}) = \mathbf{y}$, we say the distribution is a **natural exponential family**.

Example: Bernoulli

- For the Bernoulli distribution, $y \in \{0, 1\}$, and we have

$$\begin{aligned}
\text{Ber}(y|\mu) &= \mu^y (1-\mu)^{1-y} \\
&= \exp\left[y \log \mu + (1-y) \log(1-\mu)\right] \\
&= \exp\left[y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right] \\
&= h(y) \exp\left[\eta^\top \mathcal{T}(y) - A(\eta)\right],
\end{aligned}$$

where

- $\eta = \log\left(\frac{\mu}{1-\mu}\right)$, known as the **log-odds ratio**
- $\mathcal{T}(y) = y$
- $A(\eta) = -\log(1-\mu) = \log(1+e^\eta)$
- $h(y) = 1$

Example: Univariate Normal Distribution

- The univariate Normal distribution can be written as

$$\begin{aligned}\mathcal{N}(y|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left[-\frac{1}{2\sigma^2}y^2+\frac{\mu}{\sigma^2}y-\frac{1}{2\sigma^2}\mu^2\right] \\ &= \frac{1}{Z(\boldsymbol{\eta})}\exp\left(\boldsymbol{\eta}^\top\boldsymbol{\mathcal{T}}(y)\right),\end{aligned}$$

where

- $\boldsymbol{\eta} = \begin{bmatrix}\mu/\sigma^2 \\ -1/2\sigma^2\end{bmatrix}$
- $\boldsymbol{\mathcal{T}}(y) = \begin{bmatrix}y \\ y^2\end{bmatrix}$
- $Z(\boldsymbol{\eta}) = \sqrt{2\pi}\sigma\exp\left\{\frac{\mu^2}{2\sigma^2}\right\}$
- h(y) = 1

Example: Poisson Distribution

- This one is simpler - we can write $\text{Poi}(y|\mu) = \exp(-\mu)\frac{\mu^y}{y!}$ as

$$\text{Poi}(y|\mu) = \frac{h(y)}{Z(\eta)}\exp(\eta y),$$

where $\eta = \log \mu$, $h(y) = 1/y!$, and $Z(\eta) = \exp(\mu)$, or as

$$\text{Poi}(y|\mu) = h(y)\exp(\eta y - A(\eta)),$$

with $A(\eta) = \mu$.

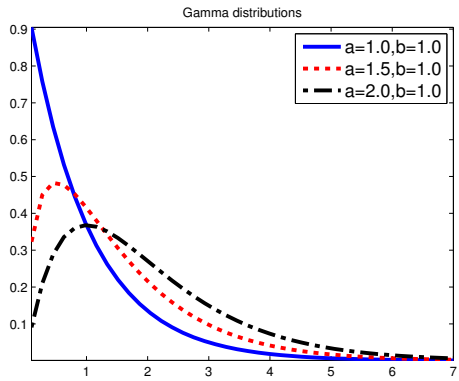
- Importantly, the mean μ can be recovered the canonical parameter η as $\mu = \exp(\eta)$.

Example: Gamma distribution (I)

- The PDF of the Gamma distribution is:

$$\mathrm{Ga}(y|a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by),$$

where $a > 0$ (shape), and $b > 0$ (rate). $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$ is the Gamma function.



Example: Gamma distribution (II)

- The PDF of the Gamma distribution is:

$$\text{Ga}(y|a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by),$$

where $a > 0$ (shape), and $b > 0$ (rate). $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$ is the Gamma function.

- This is also an exponential family with
  - $\eta = \begin{bmatrix} a-1 \\ -b \end{bmatrix}$
  - $\mathcal{T}(y) = \begin{bmatrix} \log y \\ y \end{bmatrix}$
  - $Z(\eta) = \frac{\Gamma(a)}{b^a}$
  - $h(y) = 1$

Why should we care about the exponential family?

- The exponential family are the only distributions with finite-sized sufficient statistics.
  - For a data set D of any size, there is a fixed size statistic $\mathcal{T}(D)$ that captures all the information D provides about the parameter η of the distribution.
  - This result is known as the **Pitman–Koopman–Darmois theorem**.

- The exponential family is the only family of distributions for which conjugate priors exist.
  - This makes **Bayesian inference** much simpler.

- Given some constraints on an otherwise unknown probability distribution, the distribution that satisfies these constraints while making the minimal set of additional assumptions will be a member of the exponential family.

- **Why we care today:** the exponential family is the basis **generalised linear models**.

Contents

Generalised Linear Models (I)

- Assume that the output distribution $p(y|\eta)$ is an exponential family.
- Define η as a function of $\mathbf{w}^\top \mathbf{x}$ - model the relationship between the inputs $\mathbf{x}$ and the output y.
- Linear, logistic, and Poisson regression are examples of generalised linear models (GLMs).

Generalised Linear Models (II)

- First consider another way of writing an *unconditional* exponential family over y.
- Replace η with two parameters θ and σ^2, and let $\mathcal{T}(y) = y$. We then write

$$p(y|\theta, \sigma^2) = \exp\left[\frac{y\theta - A(\theta)}{\sigma^2} + c(y, \sigma^2)\right]$$

- Here σ^2 is the **dispersion parameter**, while we now call θ the natural parameter. $c(y, \sigma^2)$ is a normalization term, independent of θ.
- θ is unknown, and dependent on $\mathbf{x}$, while σ^2 is known and fixed.

Generalised Linear Models (III)

- θ depends on the mean $\mu = \mathbb{E}[y]$ of the distribution
- For example, if $p(y|\theta, \sigma^2)$, θ is the log-odds ratio

$$\theta = \log\left(\frac{\mu}{1-\mu}\right).$$

- Let ψ denote the function mapping from the mean to the natural parameter, that is $\theta = \psi(\mu)$.
- ψ is uniquely determined by the form of the specific exponential family.
- Furthermore, the mapping ψ is invertible, such that $\mu = \psi^{-1}(\theta)$.

Link Functions

- The mean μ depends on $\mathbf{x}$ through a **link function** g as

$$\mu = g^{-1}(\mathbf{w}^\top \mathbf{x}).$$

- By convention we call g the link function and g^{-1} the **mean function**.

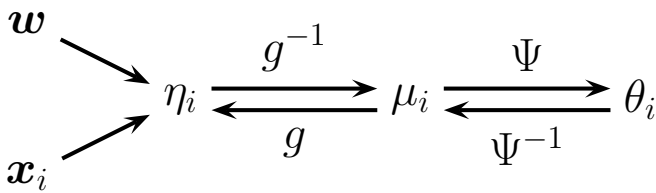
- We can choose any g we like, so long as it is invertible and g^{-1} has the same range as μ.

- For example, in logistic regression we choose $\mu = g^{-1}(\mathbf{w}^\top \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x})$, that is

$$g(s) = \log \frac{s}{1-s}$$

Relationships between functions

Let $\eta_i = \mathbf{w}^\top x_i$ be the output of the linear model for x_i, then



$g^{-1}()$ is the **mean function**. $g()$ is the **link function**. (Figure credit: Murphy 2012)

Canonical Link Functions

- A useful special case is where $g = \psi$.
- In this case we have $\theta = \mathbf{w}^\top \mathbf{x}$, so the conditional distribution on y becomes

$$p(y|\mathbf{x}, \mathbf{w}, \sigma^2) = \exp\left[\frac{y\mathbf{w}^\top \mathbf{x} - A(\mathbf{w}^\top \mathbf{x})}{\sigma^2} + c(y, \sigma^2)\right].$$

- This is called the **canonical link function**.

Example: Linear Regression

- In linear regression, the response variable y is normally distributed,

$$\begin{aligned} p(y|\mu,\sigma^2) &= \mathcal{N}(y|\mu,\sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \\ &= \exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} + \log\left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\right)\right] \\ &= \exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right]. \end{aligned}$$

- Here the link function is simple the identity, such that $\theta = \mu = \mathbf{w}^\top \mathbf{x}$.

- With $A(\theta) = \mu^2/2$, $\mathbb{E}[y] = \mu$, and $\mathrm{Var}[y] = \sigma^2$.

Example: Logistic Regression

- In logistic regression, the response variable y follows a Bernoulli distribution

$$\begin{aligned} p(y|\mu, \sigma^2) &= \mu^y (1-\mu)^y \\ &= \exp\left[\log\left(\frac{\mu}{1-\mu}\right) y - (-\log(1-\mu))\right]. \end{aligned}$$

- The link function is the logit function, such that $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \mathbf{w}^\top \mathbf{x} = \theta$.

- With $A(\theta) = -\log(1-\sigma(\theta))$, $\mathbb{E}[y] = \sigma(\theta)$, and $\mathrm{Var}[y] = \sigma(\theta)(1-\sigma(\theta))$.

Example: Poisson regression

- In Poisson regression, the response variable y follows a Poisson distribution

$$p(y|\mathbf{x}, \mathbf{w}, \sigma^2) = \exp[y \log(\mu) - \mu - \log(y!)].$$

- The link function is $g(\mu) = \log \mu = \mathbf{w}^\top \mathbf{x} = \theta$.

- With $A(\theta) = \exp(\theta)$, $\mathbb{E}[y] = \exp(\theta)$.

Canonical Link Functions

Distribution	Link Function $g(\mu)$	$\theta = \psi(\mu)$	$\mu = \psi^{-1}(\theta)$
$\mathcal{N}(\mu, \sigma^2)$	identity	$\theta = \mu$	$\mu = \theta$
Ber(μ)	logit	$\theta = \log\left(\frac{\mu}{1-\mu}\right)$	$\mu = \sigma(\theta)$
Poi(μ)	logarithm	$\theta = \log(\mu)$	$\mu = \exp(\theta)$
Ga(a, b)	inverse	$\theta = \mu^{-1}$	$\mu = \theta^{-1}$.

Mean and Variance of the Response Variable

- Recall that we referred to A as the *cumulant function*.
- It can be shown that

$$\mathbb{E}[y|\mathbf{x}, \mathbf{w}, \sigma^2] = \mu = A'(\theta)$$
$$\text{Var}[y|\mathbf{x}, \mathbf{w}, \sigma^2] = A''(\theta)\sigma^2.$$

  that is, $A(s + \theta)$ is the **cumulant generating function** for y.

- Note that the dispersion parameter σ^2 is not the variance of y - which may depend on $\mathbf{x}$.

Contents

Log-likelihood for a GLM

- One of the appealing properties of GLMs is that they can be fit using exactly the same methods that we used for logistic regression.

- Again, let $\eta_i = \mathbf{w}^\top x_i$. The log-likelihood of D has the following form

$$\mathcal{L}(D; \mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{\sigma^2} \sum_{i=1}^{N} [\eta_i y_i - A(\eta_i)].$$

Gradient of the Log-Likelihood

- Let $\ell_i = \log p(y_i|\mathbf{x}_i, \mathbf{w})$. We can find the gradient of ℓ_i using the chain rule:

$$\begin{aligned}\frac{\partial \ell_i}{\partial w_j} &= \frac{d\ell_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial w_j} \\ &= (y_i - A'(\theta_i))\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}x_{i,j} \\ &= (y_i - \mu_i)\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}x_{i,j}.\end{aligned}$$

- If we use a canonical link function such that $\theta_i = \eta_i$, this simplifies to

$$\mathbf{g}(\mathbf{w}) = \frac{1}{\sigma^2}\left[\sum_{i=1}^{N}(y_i - \mu_i)\mathbf{x}_i\right] = \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{y} - \boldsymbol{\mu}),$$

where $\mathbf{X} = [x_1, \cdots, x_N]^\top$ and $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_N]^\top$.

Hessian of the Log-Likelihood

- For faster convergence, we typically use a second-order methods (e.g., Newton's method).
- Again, if we use a canonical link function, the Hessian matrix is given by

$$\mathbf{H}(\mathbf{w}) = -\frac{1}{\sigma^2}\sum_{i=1}^{N}\frac{d\mu_i}{d\theta_i}\mathbf{x}_i\mathbf{x}_i^\top = -\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{\Sigma}\mathbf{X},$$

  where $\mathbf{\Sigma} = \text{diag}\left(\frac{d\mu_1}{d\theta_1}, \cdots, \frac{d\mu_N}{d\theta_N}\right)$.

- This form for $\mathbf{H}(\mathbf{w})$ and $\mathbf{g}(\mathbf{w})$ allows us to use Iteratively Reweighted Least Squares (IRLS).

Iteratively Reweighted Least Squares (IRLS) Algorithm

- IRLS is really just Newton's iteration $\mathbf{w}_{k+1} = \mathbf{w}_k - \mathbf{H}_k^{-1}\mathbf{g}_k$.
- In the case maximizing the log-likelihodd of a GLM, the iterates $\mathbf{w}_k$ of Newton's method are solutions to a series of **weighted least squares** problems.
- Given $\mathbf{w}_k$, we can compute $\mathbf{w}_{k+1}$ using *any* solver for weighted least squares.

Least Squares Problems

- Remember that a least squares (LS) problem refers to

$$LS(\mathbf{w}) = \min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2,$$

  for a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N = \{\mathbf{X}, \mathbf{y}\}$.

- It can be shown that the vector $\mathbf{w}$ minimizing $LS(\mathbf{w})$ is given by

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Weighted Least Squares Problems

- A weighted least squares (*WLS*) problem refers to

$$WLS(\mathbf{w}) = \min_{\mathbf{w}} \sum_{i=1}^{N} r_i (y_i - \mathbf{w}^\top \mathbf{x}_i)^2,$$

for a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i, r_i)\}_{i=1}^N = \{\mathbf{X}, \mathbf{R}, \mathbf{y}\}$, with $\mathbf{R} = \text{diag}(r_1, \ldots, r_N)$.

- It can be shown that the vector $\mathbf{w}$ that minimises $WLS(\mathbf{w})$ is given as

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{R} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{R} \mathbf{y}.$$

Iterative Reweighted Least Squares

- Newton's method for the log-likelihood of the GLM follows as

$$\begin{aligned}\mathbf{w}_{k+1} &= \mathbf{w}_k - \mathbf{H}_k^{-1}\mathbf{g}_k \\ &= \mathbf{w}_k + (\mathbf{X}^\top \Sigma_k \mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{y} - \boldsymbol{\mu}_k) \\ &= (\mathbf{X}^\top \Sigma_k \mathbf{X})^{-1}[\mathbf{X}^\top \Sigma_k \mathbf{X}\mathbf{w}_k + \mathbf{X}^\top(\mathbf{y} - \boldsymbol{\mu}_k)] \\ &= (\mathbf{X}^\top \Sigma_k \mathbf{X})^{-1}\mathbf{X}^\top \Sigma_k \mathbf{z}_k,\end{aligned}$$

  where $\mathbf{z}_k = \mathbf{X}\mathbf{w}_k + \Sigma_k^{-1}(\mathbf{y} - \boldsymbol{\mu}_k)$ is known as the **working response**.

- At iteration k, we can see that $\mathbf{w}_{k+1}$ is the solution to the weighted least squares problem

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{d\mu_i}{d\theta_i}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2$$

  where we replace the diagonal weight matrix $\mathbf{R}$ with Σ_k, which changes at each iteration.

Contents

GLM available in Spark

- Spark's implementation of GLMs supports the following families and link functions:

Family	Response type	Supported links
Gaussian	Continuous	Identity*, Log, Inverse
Binomial	Binary	Logit*, Probit, CLogLog
Poisson	Count	Log*, Identity, Sqrt
Gamma	Continuous	Inverse*, Identity, Log
Tweedie	Zero-inflated continuous	Power link function

where * stands for canonical link.

- For a random variable Z that obeys a Tweedie distribution, the variance $\mathrm{var}(Z)$ relates to the mean $\mathbb{E}(Z)$ by the power law: $\mathrm{var}(Z) = a\mathbb{E}(Z)^p$, where a and p are positive constants.

- Set using `family` and `link` arguments to `GeneralizedLinearRegression()`.

`GeneralizedLinearRegression()`

- It uses IRLS (Iterative Reweighted Least Squares) for optimisation.

- **Limitations:**
  - Only supports ℓ_2 regularisation.
  - Spark currently only supports up to 4096 features through its `GeneralizedLinearRegression` interface.
  - Will throw an exception if this constraint is exceeded.

Parameters to Adjust

- **`maxIter`**: max number of iterations.
- **`regParam`**: regularization parameter (≥ 0).
- **`family`**: name of exponential family used in the model.
- **`link`**: name of link function, which defines the relationship between the linear predictor $\mathbf{w}^\top \mathbf{x}$ and the mean of the output distribution.

Special cases: `LinearRegression()`

- If your family is “gaussian” and your link function is the “identity”, your model will be equivalent to linear regression.
- Recommended that you use `LinearRegression()` instead of `GeneralizedLinearRegression()` when this is the case.
- `LinearRegression()` allows for ℓ_2, ℓ_1 and elastic net regularization (fit using L-BFGS or OWL-QN).

Special cases: `LogisticRegression()`

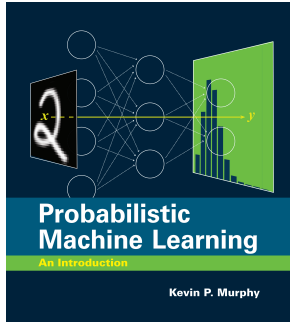
- If your family is “binomial” and your link function is “logit”, your model will be equivalent to logistic regression.
- Recommended that you use `LogisticRegression()` instead of `GeneralizedLinearRegression()` when this is the case.
- `LogisticRegression()` allows for ℓ_2, ℓ_1 and elastic net regularization (fit using L-BFGS or OWL-QN).

Contents

References and Recommended Reading

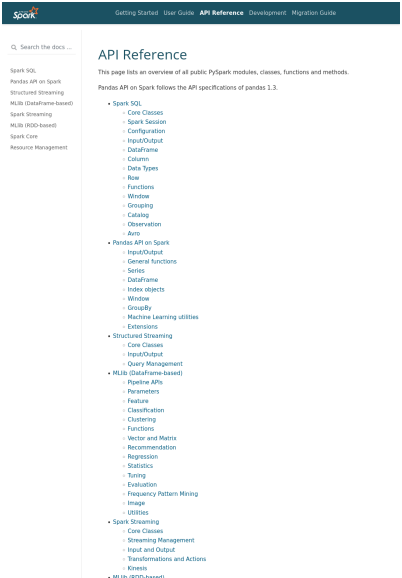
Book: *Probabilistic Machine Learning: An Introduction* by Kevin P Murphy, 2022.

- Section 3.4 The exponential family, pp. 93 - 96.
- Chapter 12 Generalized linear models, pp. 413 - 419.



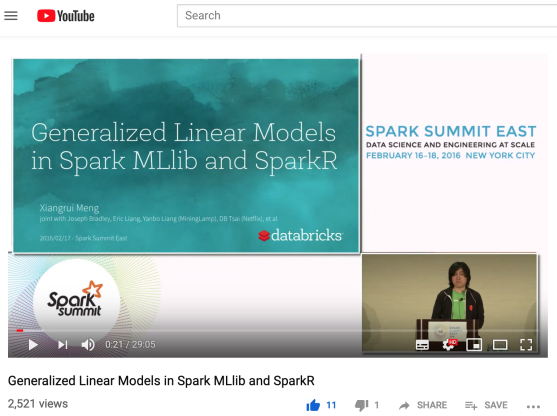
References and Recommended Reading

Website: *Spark Python API Documentation*.



References and Recommended Reading

Youtube video: *Generalized Linear Models in Spark MLlib and SparkR* by Xiangrui Meng (from Databricks).



https://www.youtube.com/watch?v=PSZW6hcQ_7w

References and Recommended Reading

Paper: *Map-Reduce for Machine Learning on Multicore* by C-T Chu et al. (2006).

Map-Reduce for Machine Learning on Multicore

Cheng-Tao Chu * chengtao@stanford.edu

Sang Kyun Kim * skkim38@stanford.edu

Yi-An Lin * ianl@stanford.edu

YuanYuan Yu * yuanyuan@stanford.edu

Gary Bradski *† garybradski@gmail

Andrew Y. Ng * ang@cs.stanford.edu

Kunle Olukotun * kunle@cs.stanford.edu

* CS. Department, Stanford University 353 Serra Mall, Stanford University, Stanford CA 94305-9025.

† Rexee Inc.

Abstract

We are at the beginning of the multicore era. Computers will have increasingly many cores (processors), but there is still no good programming framework for these architectures, and thus no simple and unified way for machine learning to take advantage of the potential speed up. In this paper, we develop a broadly applicable parallel programming method, one that is easily applied to *many* different learning algorithms. Our work is in distinct contrast to the tradition in machine learning of designing (often ingenious) ways to speed up a *single* algorithm at a time. Specifically, we show that algorithms that fit the Statistical Query model [15] can be written in a certain "summation form," which allows them to be easily parallelized on multicore computers. We adapt Google's map-reduce [7] paradigm to demonstrate this parallel speed up technique on a variety of learning algorithms including locally weighted linear regression (LWLR), k-means, logistic regression (LR), naive Bayes (NB), SVM, ICA, PCA, gaussian discriminant analysis (GDA), EM, and backpropagation (NN). Our experimental results show basically linear speedup with an increasing number of processors.