Computational Statistics
Linear Classification

Luca Bortolussi

Department of Mathematics and Geosciences
University of Trieste

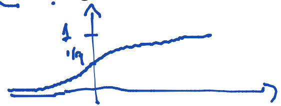
Office 238, third floor, H2bis
luca@dmi.units.it

Trieste, Winter Semester 2015/2016

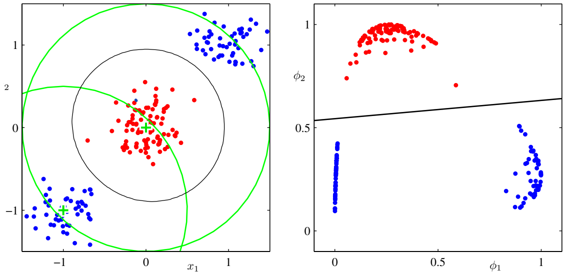
Outline

LOGIT AND PROBIT REGRESSION (BINARY CASE)

- We model directly the conditional class probabilities $p(\mathcal{C}_1|\mathbf{x}) = \boxed{f(\mathbf{w}^T\phi(\mathbf{x}))}$ after a (nonlinear) mapping of the features $\phi(\mathbf{x}) = \phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})$.
- Common choices for f are the logistic or logit function $\boxed{\sigma(a) = \frac{1}{1+e^{-a}}}$ and the probit function $\psi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta$.



- We will focus on logistic regression.
- The non-linear embedding is an important step



LOGISTIC REGRESSION

$p(C_1|x) = p(C_1|\phi(x)) =$

- We assume $\boxed{p(C_1|\phi)} = y(\phi) = \boxed{\sigma(\mathbf{w}^T\phi)}$ where $\phi = \phi(\mathbf{x})$ and $\phi_i = \phi(\mathbf{x_i})$.
- As $y = y(\phi(\mathbf{x})) \in [0, 1]$ we interpret is as the probability of assigning input $\mathbf{x}$ to class 1, so that the likelihood is

$x_i \quad t_i \in \{0, 1\}$

$y_i = p(C_1|\phi(x_i)) = \sigma(w^T\phi(x_i))$

$$p(\mathbf{t}|\mathbf{w}) = \prod_{i=1}^{N} y_i^{t_i}(1 - y_i)^{1-t_i}$$

$y_i \cdot 1$ set $t_i = 1$

$1 \cdot (1 - y_i)$ set $t_i = 0$

where $y_i = \sigma(\mathbf{w}^T\phi_i)$.

- We need to minimise minus the log-likelihood, i.e.

$$E(\mathbf{w}) = -\log p(\mathbf{t}|\mathbf{w}) = -\sum_{i=1}^{N} t_i \log y_i + (1 - t_i)\log(1 - y_i)$$

$y_i = y_i(w)$

$\sigma'(w^T\phi(x_i)) = \frac{d\sigma}{dz}\Big| \cdot \phi(x_i)$

NUMERICAL OPTIMISATION

$\frac{d}{da}\sigma(a) = \sigma(a)(1-\sigma(a))$

- The gradient of $E(\mathbf{w})$ is $\nabla E(\mathbf{w}) = \sum_{i=1}^{N}(y_i - t_i)\phi_i$. The equation $\boxed{\nabla E(\mathbf{w}) = 0}$ has no closed form solution, so we need to solve it numerically.
- One possibility is gradient descend. We initialise $\mathbf{w}^0$ to any value and then update it by

$$\mathbf{w}^{n+1} = \mathbf{w}^n + \eta \nabla E(\mathbf{w}^n)$$

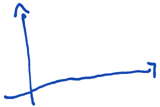
where the method converges for η small.

- We can also use stochastic gradient descent for online training, using the update rule for $\mathbf{w}$:

$$\mathbf{w}^{n+1} = \mathbf{w}^n + \eta \nabla_{n+1} E(\mathbf{w}^n),$$

with $\boxed{\nabla_n E(\mathbf{w})} = (y_n - t_n)\phi_n$

Logistic regression: overfitting



$$\sigma(w^T\phi(x)) = \frac{1}{2} \Leftrightarrow \boxed{w^T\phi(x) = 0}$$

- If we allocate each point $\mathbf{x}$ to the class with highest probability, i.e. maximising $\sigma(\mathbf{w}^T\phi(\mathbf{x}))$, then the separating surface is an hyperplane in the feature space and is given by the equation $\mathbf{w}^T\phi(\mathbf{x}) = 0$.
- If the data is linearly separable in the feature space, then any separable hyperplane is a solution, and the magnitude of $\mathbf{w}$ tends to go to infinity during optimisation. In this case, the logistic function converges to the Heaviside function.
- To avoid this issue, we can add a regularisation term to $E(\mathbf{w})$, thus minimising $E(\mathbf{w}) + \boxed{\alpha\mathbf{w}^T\mathbf{w}}$.

NEWTON-RAPSON METHOD

- As an alternative optimisation, we can use the Newton-Rapson method, which has better convergence properties.
- The update rule reads:

$$\mathbf{w}^{new} = \mathbf{w}^{old} - \mathbf{H}^{-1}\nabla E(\mathbf{w}^{old})$$

  where $\mathbf{H}$ is the Hessian of $E(\mathbf{w})$.
- For logistic regression, we have $\nabla E(\mathbf{w}) = \Phi^T(\mathbf{y} - \mathbf{t})$ and $\mathbf{H} = \Phi^T \mathbf{R} \Phi$, with R diagonal matrix with elements $R_{nn} = y_n(1 - y_n)$.
- It is easy to check that the Hessian is positive definite, hence the function $E(\mathbf{w})$ is convex and has a unique minimum.

MULTI-CLASS LOGISTIC REGRESSION

- We can model directly the multiclass conditional probability, using the soft-max function:

$\phi(x)$

$\forall k = 1, \dots, K$

$w_k^T \cdot \phi(x)$

$$k = 1, \dots \rightarrow \boxed{p(C_k|\mathbf{x})} = y_k(\mathbf{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

with $\boxed{a_k = \mathbf{w_k}\phi(\mathbf{x})}$. It holds $\frac{\partial y_k(\mathbf{x})}{\partial a_j} = y_k(\delta_{kj} - y_j)$

- Using the boolean encoding of the outputs, the likelihood is

$t_n = (t_{n1}, \dots, t_{nk}, \dots, t_{nK})$

$$p(\mathbf{T}|\mathbf{w_1}, \dots, \mathbf{w_K}) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(C_k|\phi_n)^{t_{nk}} = y_{nk}^{t_{nk}} \quad \prod_{n,K} y_{nk}^{t_{nk}}$$

- Hence we need to minimise

$$\boxed{E(\mathbf{w_1}, \dots, \mathbf{w_K}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk}}$$

MULTI-CLASS LOGISTIC REGRESSION

- $E(\mathbf{w_1}, \ldots, \mathbf{w_K})$ has gradient

$$\nabla_{\mathbf{w_j}} E(\mathbf{w_1}, \ldots, \mathbf{w_K}) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\phi_n$$

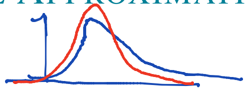
- and Hessian with blocks given by

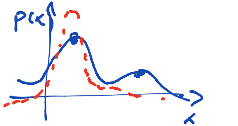
$$\nabla_{\mathbf{w_k}} \nabla_{\mathbf{w_j}} E(\mathbf{w_1}, \ldots, \mathbf{w_K}) = -\sum_{n=1}^{N} y_{nk}(I_{kj} - y_{nj})\phi_n \phi_n^T$$

- Also in this case the Hessian is positive definite, and we can use the Newton-Rapson algorithm for optimisation

Outline

LAPLACE APPROXIMATION - 1 DIMENSION





- It is a general technique to locally approximate a general distribution around a mode with a Gaussian.
- Consider a 1d distribution $p(z) = \boxed{\frac{1}{Z} f(z)}$ where $Z = \int f(z)dz$ is the normalisation constant.

  $f(z) \geq 0$
  Int.
  $Z = \int_{-\infty}^{\infty} f(z)dz$
- Pick a mode z_0 of $f(z)$, i.e. a point such that $\frac{d}{dz} f(z_0) = 0$.
- As the logarithm of the Gaussian density is quadratic, we consider a Taylor expansion of $\boxed{\log f(z)}$ around z_0:

$$\log f(z) \approx \log f(z_0) - \frac{1}{2} A (z - z_0)^2$$

with $A = \boxed{-\frac{d^2}{dz^2} \log f(z_0)}$

LAPLACE APPROXIMATION - 1 DIMENSION

- Hence we have $f(z) \approx f(z_0) \exp(-\frac{1}{2}A(z - z_0)^2)$. Now, we seek the best Gaussian $q(z)$ approximating $p(z)$ around the mode z_0, requiring $A > 0$. This is clearly given by

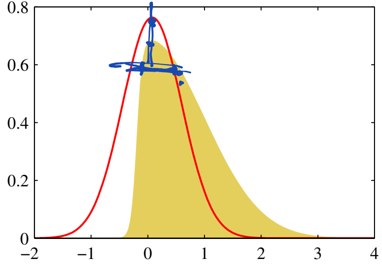
$$q(z) = \left(\frac{A}{2\pi}\right)^{\frac{1}{2}} \exp(-\frac{1}{2}A(z - z_0)^2)$$

- We also have that $Z \approx f(z_0)\left(\frac{A}{2\pi}\right)^{-\frac{1}{2}}$

$\int f(z)\,dz =$

$\int f(z_0) \cdot C \cdot q(z)\,dz =$

$f(z_0) \cdot C$



LAPLACE APPROXIMATION - N DIMENSION

- In n dimensions, we proceed in the same way. Given a density $p(\mathbf{z}) = \frac{1}{Z} f(\mathbf{z})$, we find a mode $\mathbf{z_0}$ (so that $\nabla \log f(\mathbf{z_0}) = \mathbf{0}$, and approximate $\log f(\mathbf{z})$ around $\mathbf{z_0}$ by Taylor expansion, obtaining

$$\log f(\mathbf{z}) = \log f(\mathbf{z_0}) - \frac{1}{2}(\mathbf{z} - \mathbf{z_0})^T \mathbf{A}(\mathbf{z} - \mathbf{z_0})$$

where $\boxed{\mathbf{A} = -\nabla\nabla \log f(\mathbf{z_0})}$.

- This gives a Gaussian approximation around $\mathbf{z_0}$ by

$$q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{z_0}, \mathbf{A}^{-1})$$

- Furthermore $\boxed{Z \approx \frac{(2\pi)^{n/2}}{|\mathbf{A}|^{1/2}} f(\mathbf{z_0})}$

MODEL COMPARISON

- We can use Laplace approximation for the marginal likelihood in a model comparison framework.
- Consider data $\mathcal{D}$ and a model $\mathcal{M}$ depending on parameters θ. We fix a prior $\mathcal{P}(\theta)$ over θ and compute the posterior by Bayes theorem:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

- Here $p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$ is the marginal likelihood. It fits in the previous framework by setting $Z = p(\mathcal{D})$, and
$f(\theta)$ $f = p(\mathcal{D}|\theta)p(\theta)$.

BIC

- By Laplace approximation around the maximum a-posteriori estimate θ_{MAP}:

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\theta_{MAP}) + \log p(\theta_{MAP}) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{A}|$$

where $\mathbf{A} = -\nabla\nabla p(\mathcal{D}|\theta_{MAP})p(\theta_{MAP})$. The last three terms in the sum penalise the log likelihood in terms of model complexity.

- A crude approximation of them is

$$logp(\mathcal{D}) \approx \log p(\mathcal{D}|\theta_{MAP}) - \frac{1}{2}M\log N$$

which is known as Bayesian Information Content, and can be used to penalise log likelihood w.r.t. model complexity, to compare different models.

Outline

THE BAYESIAN WAY

$p(w) = \mathcal{N}(0, \alpha^{-1} I)$

- To recast logistic regression in a Bayesian framework, we need to put a prior on $p(\mathbf{w})$ of the coefficients $\mathbf{w}$ of $\sigma(\mathbf{w}^T\phi(\mathbf{x}))$ and compute the posterior distribution on $\mathbf{w}$ by Bayes theorem. Then we can make predictions by integrating out the parameters.
- Assume a Gaussian prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m_0}, \mathbf{S_0})$. The posterior is $p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w})p(\mathbf{t}|\mathbf{w})$, and the log-posterior is

$$\log p(\mathbf{w}|\mathbf{t}) = -\frac{1}{2}(\mathbf{w}-\mathbf{m_0})^T S_0^{-1}(\mathbf{w}-\mathbf{m_0}) + \sum_{i=1}^{N}[t_i \log y_i + (1-t_i)\log(1-y_i)] + c$$

where $y_i = \sigma(\mathbf{w}\phi(\mathbf{x_i}))$.

- Computing the marginal likelihood and the normalisation constant is analytically intractable, due to quadratic and logistic terms. Hence we do a Laplace approximation of the posterior.

LAPLACE APPROXIMATION OF THE POSTERIOR

- Given $\log p(\mathbf{w}|\mathbf{t})$, we first find the maximum a-posteriori $\mathbf{w_{MAP}}$, by running a numerical optimisation, and then obtain the Laplace approximation computing the Hessian matrix at $\mathbf{w_{MAP}}$ and inverting it, obtaining

$$\mathbf{S_N} = -\nabla\nabla \log \boxed{p(\mathbf{w}|\mathbf{t})} = \mathbf{S_0}^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\phi(\mathbf{x_n})\phi(\mathbf{x_n})^T$$

evaluated at $\mathbf{w} = \mathbf{w_{MAP}}$.

- Hence, the Laplace approximation of the posterior is

$$\boxed{q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w_{MAP}}, \mathbf{S_N})}$$

PREDICTIVE DISTRIBUTION

$x - w^T\phi(x) \to \sigma(w^T\phi(x))$

- The predictive distribution for class C_1 is given by

$$p(C_1|\phi, \mathbf{t}) = \int p(C_1|\phi, \mathbf{w}, \mathbf{t})q(\mathbf{w})d\mathbf{w} = \int \sigma(\mathbf{w}^T\phi(\mathbf{x}))q(\mathbf{w})d\mathbf{w}$$

- This multi-dimensional integral can be simplified by noting that it depends on $\mathbf{w}$ only on the 1-dim projection $a = \mathbf{w}^T\phi(\mathbf{x})$, and that q restricted to this direction is still a Gaussian distribution $q(a)$ with mean and variance

$$\mu_a = \mathbf{w_{MAP}}^T\phi(\mathbf{x}) \quad \sigma_a^2 = \phi(\mathbf{x})^T\mathbf{S_N}\phi(\mathbf{x})$$

- Hence we have

$$p(C_1|\phi, \mathbf{t}) = \int \sigma(a)q(a)da$$

PROBIT APPROXIMATION

$\Psi(a) = \int_{-\infty}^{a} \mathcal{N}(z \mid 0,1)\,dz$

$\sigma(a) \approx \Psi(\lambda a)$

- The integral $p(C_1|\phi, \mathbf{t}) = \int \sigma(a)q(a)da$ can be approximated by approximating the logistic function by the probit: $\sigma(a) = \Psi(\lambda a)$, where λ is obtained by matching derivatives at zero and is $\boxed{\lambda^2 = \pi/8}$.
- We then use

$$\int \Psi(\lambda a)\mathcal{N}(a|\mu, \sigma^2) = \boxed{\Psi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right)}$$

and approximate back to the logistic to get

$$\boxed{p(C_1|\phi, \mathbf{t}) \approx \sigma(\kappa(\sigma_a^2)\mu_a)}$$

with $\kappa(\sigma_a^2) = \underbrace{(1 + \pi\sigma_a^2/8)^{-1/2}}$