TDA231
Logistic regression

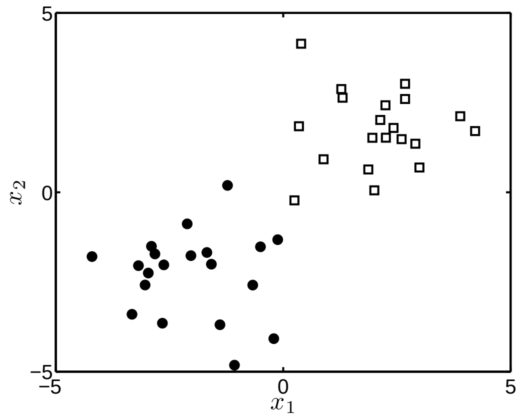
Devdatt Dubhashi
dubhashi@chalmers.se

Dept. of Computer Science and Engg.
Chalmers University

February 19, 2016

Introduction
D. Dubhashi
Logistic regression
Point estimate
MCMC sampling

Some data



Introduction
D. Dubhashi
Logistic regression
Point estimate
MCMC sampling

Logistic regression

- In the Bayes classifier, we built a model of each class and then used Bayes rule:

$$P(T_{\text{new}} = k | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = \frac{p(\mathbf{x}_{\text{new}} | T_{\text{new}} = k, \mathbf{X}, \mathbf{t}) P(T_{\text{new}} = k)}{\sum_j p(\mathbf{x}_{\text{new}} | t_{\text{new}} = j, \mathbf{X}, \mathbf{t}) P(t_{\text{new}} = j)}$$

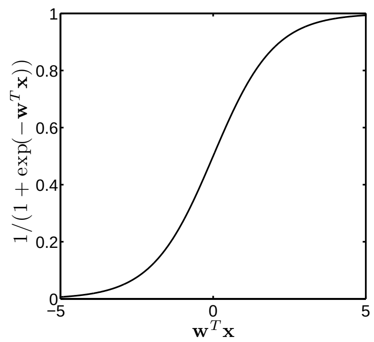
- Alternative is to directly model $P(T_{\text{new}} = k | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = f(\mathbf{x}_{\text{new}}; \mathbf{w})$ with some parameters $\mathbf{w}$.
- We've seen $f(\mathbf{x}_{\text{new}}; \mathbf{w}) = \mathbf{w}^{\mathsf{T}} \mathbf{x}_{\text{new}}$ before – can we use it here?
  - No – *output is unbounded and so can't be a probability.*
- But, can use $P(T_{\text{new}} = k | \mathbf{x}_{\text{new}}, \mathbf{w}) = h(f(\mathbf{x}_{\text{new}}; \mathbf{w}))$ where $h(\cdot)$ *squashes* $f(\mathbf{x}_{\text{new}}; \mathbf{w})$ to lie between 0 and 1 – a probability.

Introduction
D. Dubhashi
Logistic regression
Point estimate
MCMC sampling

$h(\cdot)$

- For logistic regression (binary), we use the sigmoid function:

$$P(T_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{w}) = h(\mathbf{w}^{\mathsf{T}} \mathbf{x}_{\text{new}}) = \frac{1}{1 + \exp(-\mathbf{w}^{\mathsf{T}} \mathbf{x}_{\text{new}})}$$



Bayesian logistic regression

- Recall Bayesian ideas ....
- In theory, if we place a *prior* on $\mathbf{w}$ and define a *likelihood* we can obtain a *posterior*:

$$p(\mathbf{w}|\mathbf{X},\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}$$

- And we can make predictions by taking expectations (averaging over $\mathbf{w}$):

$$P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}},\mathbf{X},\mathbf{t}) = \mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{t})}\left\{P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}},\mathbf{w})\right\}$$

- Sounds good so far....

Defining a prior

- Choose a Gaussian prior:

$$p(\mathbf{w}) = \prod_{d=1}^{D} \mathcal{N}(0,\sigma^2).$$

  - Prior choice is *always* important from a data analysis point of view.
  - Previously, it was also important 'for the maths'.
  - This isn't the case today – could choose any prior – no prior makes the maths easier!

Defining a likelihood

- First assume independence:

$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w})$$

- We have already defined this – it's our squashing function! If $t_n = 1$:

$$P(t_n = 1|\mathbf{x}_n,\mathbf{w}) = \frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}$$

- and if $t_n = 0$:

$$P(t_n = 0|\mathbf{x}_n,\mathbf{w}) = 1 - P(t_n = 1|\mathbf{x},\mathbf{w})$$

Posterior

$$p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2) = \frac{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)p(\mathbf{w}|\sigma^2)}{p(\mathbf{t}|\mathbf{X},\sigma^2)}$$

- Now things start going wrong.
- We can't compute $p(\mathbf{w}|\mathbf{X},\mathbf{t})$ analytically.
  - Prior is not conjugate to likelihood. No prior is!
  - This means we don't know the *form* of $p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)$
  - And we can't compute the marginal likelihood:

$$p(\mathbf{t}|\mathbf{X},\sigma^2) = \int p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)p(\mathbf{w}|\sigma^2)\; d\mathbf{w}$$

What can we compute?

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2) = \frac{p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)p(\mathbf{w}|\sigma^2)}{p(\mathbf{t}|\mathbf{X}, \sigma^2)}$$

- We can compute $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)p(\mathbf{w}|\sigma^2)$
  - Define $g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2) = p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\sigma^2)$
- Armed with this, we have three options:
  - Find the most likely value of $\mathbf{w}$ – a point estimate.
  - Approximate $p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2)$ with something easier.
  - Sample from $p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2)$.
- These examples aren't the only ways of approximating/sampling.
- They are also general techniques not unique to logistic regression.

MAP estimate

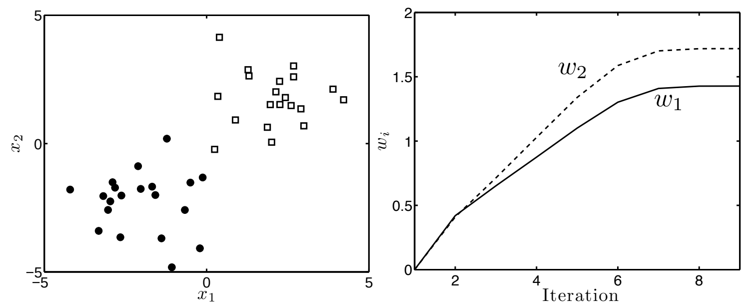
- Out first method is to find the value of $\mathbf{w}$ that maximises $p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2)$ (call it $\widehat{\mathbf{w}}$).
  - $g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2) \propto p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2)$
  - $\widehat{\mathbf{w}}$ therefore also maximises $g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$.
- Very similar to maximum likelihood but additional effect of prior.
- Known as MAP (maximum a posteriori) solution.
- Once we have $\widehat{\mathbf{w}}$, make predictions with:

$$P(t_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \widehat{\mathbf{w}}) = \frac{1}{1 + \exp(-\widehat{\mathbf{w}}^\mathsf{T}\mathbf{x}_{\text{new}})}$$

MAP

- When we met maximum likelihood, we could find $\widehat{\mathbf{w}}$ exactly with some algebra.
- Can't do that here (can't solve $\frac{\partial g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)}{\partial \mathbf{w}} = \mathbf{0}$)
- Resort to numerical optimisation:
  1. Guess $\widehat{\mathbf{w}}$
  2. Change it a bit in a way that increases $g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$
  3. Repeat until no further increase is possible.
- Many algorithms exist that differ in how they do step 2.
- e.g. Gradient Descent
  - Not covered in this course. You just need to know that sometimes we can't do things analytically and there are methods to help us! Ask John!

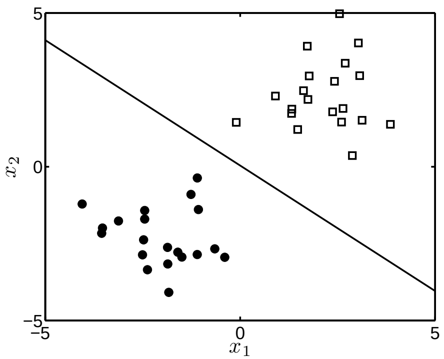
MAP – numerical optimisation for our data



- Left: Data.
- Right: Evolution of $\widehat{\mathbf{w}}$ in numerical optimisation.

Decision boundary

- Once we have $\widehat{\mathbf{w}}$, we can classify new examples.
- Decision boundary is a useful visualisation:

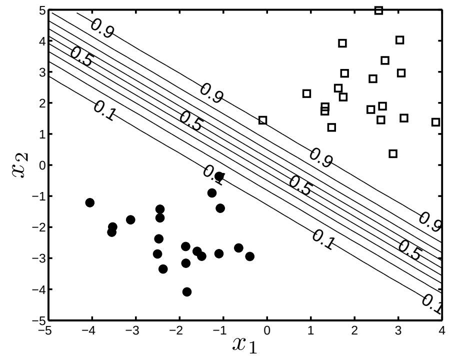


- Line corresponding to $P(T_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \widehat{\mathbf{w}}) = 0.5$.

$$0.5 = \frac{1}{2} = \frac{1}{1 + \exp(-\widehat{\mathbf{w}}^{\mathsf{T}} \mathbf{x}_{\text{new}})}.$$

So: $\exp(-\widehat{\mathbf{w}}^{\mathsf{T}} \mathbf{x}_{\text{new}}) = 1$. Or: $\widehat{\mathbf{w}}^{\mathsf{T}} \mathbf{x}_{\text{new}} = 0$

Predictive probabilities



- Contours of $P(T_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \widehat{\mathbf{w}})$.
- Do they look sensible?

Sampling feom posterior

- Suppose we can produce samples $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s, \ldots$ from $p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2)$.
- Then we can average the predictions to approximate $p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2)$:

$$\begin{aligned} P(t_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) &= \mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)} \left\{ P(t_{\text{new}} | \mathbf{x}_{\text{new}}, \mathbf{w}) \right\} \\ &\approx \frac{1}{S} \sum_{s=1}^{S} \frac{1}{1 + \exp(-\mathbf{w}_s^{\mathsf{T}} \mathbf{x}_{\text{new}})} \end{aligned}$$

MCMC sampling

- Magic! We can sample directly from $p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2)$ even though we can't compute it!
- Various algorithms exist – we'll use Metropolis-Hastings

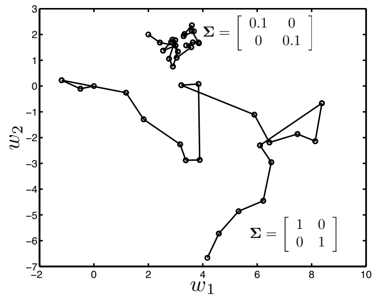
Back to the script: Metropolis-Hastings

- Produces a sequence of samples – $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s, \ldots$
- Imagine we've just produced $\mathbf{w}_{s-1}$
- MH firsts *proposes* a possible $\mathbf{w}_s$ (call it $\widetilde{\mathbf{w}_s}$) based on $\mathbf{w}_{s-1}$.
- MH then decides whether or not to *accept* $\widetilde{\mathbf{w}_s}$
  - If accepted, $\mathbf{w}_s = \widetilde{\mathbf{w}_s}$
  - If not, $\mathbf{w}_s = \mathbf{w}_{s-1}$
- Two distinct steps – proposal and acceptance.

MH – proposal

- Treat $\widetilde{\mathbf{w}_s}$ as a random variable conditioned on $\mathbf{w}_{s-1}$
- i.e. need to define $p(\widetilde{\mathbf{w}_s}|\mathbf{w}_{s-1})$
  - Note that this does not necessarily have to be similar to posterior we're trying to sample from.
- Can choose *whatever we like*!
- e.g. use a Gaussian centered on $\mathbf{w}_{s-1}$ with some covariance:

$$p(\widetilde{\mathbf{w}_s}|\mathbf{w}_{s-1}, \mathbf{\Sigma}_p) = \mathcal{N}(\mathbf{w}_{s-1}, \mathbf{\Sigma}_p)$$



MH – acceptance

- Choice of acceptance based on the following ratio:

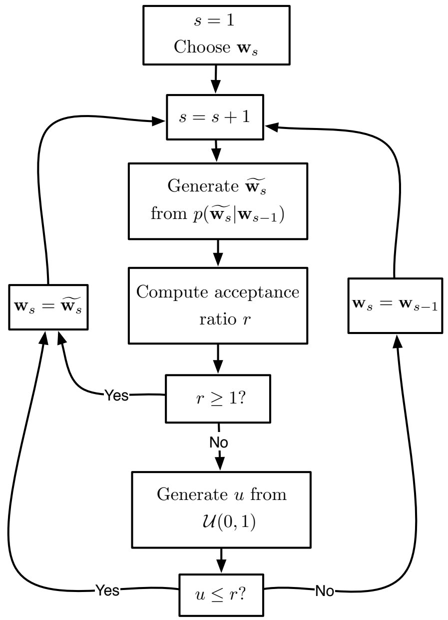
$$r = \frac{p(\widetilde{\mathbf{w}_s}|\mathbf{X}, \mathbf{t}, \sigma^2)}{p(\mathbf{w}_{s-1}|\mathbf{X}, \mathbf{t}, \sigma^2)} \frac{p(\mathbf{w}_{s-1}|\widetilde{\mathbf{w}_s}, \mathbf{\Sigma}_p)}{p(\widetilde{\mathbf{w}_s}|\mathbf{w}_{s-1}, \mathbf{\Sigma}_p)}.$$

- Which simplifies to (all of which we can compute):

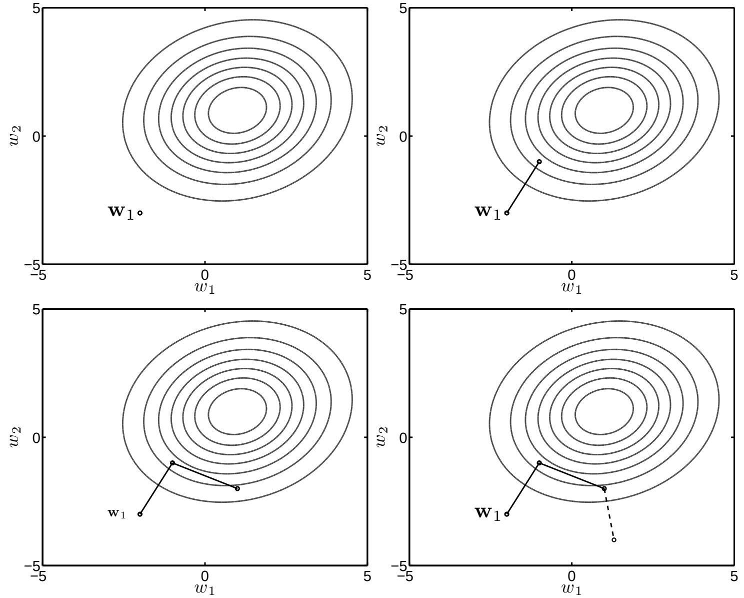
$$r = \frac{g(\widetilde{\mathbf{w}_s}; \mathbf{X}, \mathbf{t}, \sigma^2)}{g(\mathbf{w}_{s-1}; \mathbf{X}, \mathbf{t}, \sigma^2)} \frac{p(\mathbf{w}_{s-1}|\widetilde{\mathbf{w}_s}, \mathbf{\Sigma}_p)}{p(\widetilde{\mathbf{w}_s}|\mathbf{w}_{s-1}, \mathbf{\Sigma}_p)}.$$

- We now use the following rules:
  - If $r \geq 1$, accept: $\mathbf{w}_s = \widetilde{\mathbf{w}_s}$.
  - If $r < 1$, accept with probability r.
- If we do this enough, we'll eventually be sampling from $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$, no matter where we started!
  - i.e. for any $\mathbf{w}_1$

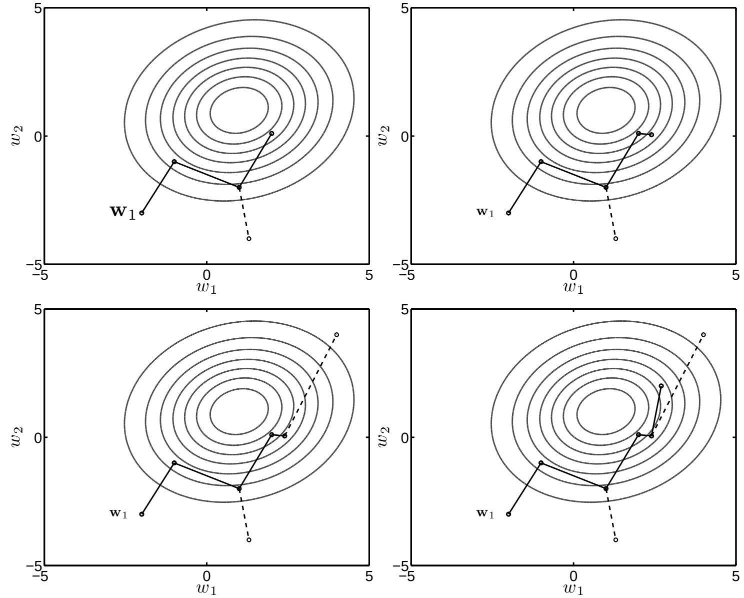
MH – flowchart



MH – walkthrough 1

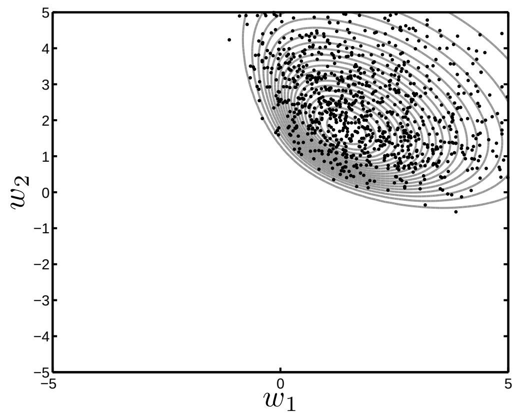


MH – walkthrough 2



Introduction
D. Dubhashi
Logistic regression
Point estimate
MCMC sampling

What do the samples look like?



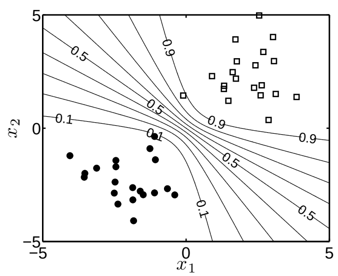
- 1000 samples from the posterior using MH.

Introduction
D. Dubhashi
Logistic regression
Point estimate
MCMC sampling

Predictions with MH

- MH provides us with a set of samples – $\mathbf{w}_1, \ldots, \mathbf{w}_S$.
- These can be used like the samples from the Laplace approximation:

$$
\begin{aligned}
P(t_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) &= \mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)} \left\{ P(t_{\text{new}} | \mathbf{x}_{\text{new}}, \mathbf{w}) \right\} \\
&\approx \frac{1}{S} \sum_{s=1}^{S} \frac{1}{1 + \exp(-\mathbf{w}_s^{\mathsf{T}} \mathbf{x}_{\text{new}})}
\end{aligned}
$$



- Contours of $P(t_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2)$

Introduction
D. Dubhashi
Logistic regression
Point estimate
MCMC sampling

Summary

- Introduced logistic regression – a probabilistic binary classifier.
- Saw that we couldn't compute the posterior.
- Introduced examples of two alternatives:
  - Point estimate – MAP solution.
  - Sample – Metropolis-Hastings.
- Second is better than the last (in terms of predictions)....
- ...but each has greater complexity!
- To think about:
  - What if posterior is multi-modal?