ISTA 421 + INFO 521
Introduction to Machine Learning

Lecture 16: Estimation I
Gradient Methods Continued
Newton-Raphson

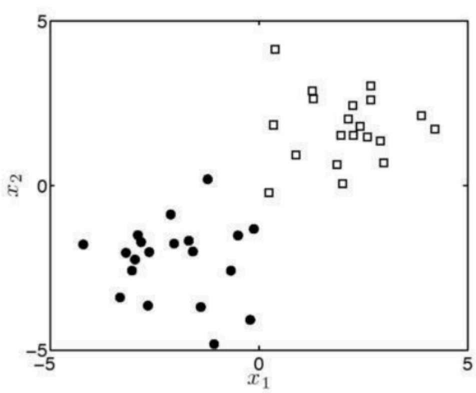
Clay Morrison
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

18 October 2017

1

Binary Classification!

- A very common type of problem
- *Many* different approaches; we'll start with a probabilistic method: logistic regression



two attributes (x_1 and x_2)

binary target, $t = \{0, 1\}$

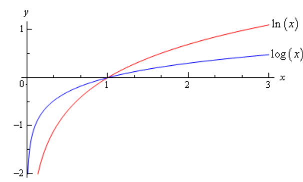
t = 0 are dark circles
t = 1 are white squares

2

The Log-odds

- The logistic function is formally derived as a result of a linear model of the ***log-odds*** (aka the ***logit***):

$$\log\left(\frac{P(T_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})}{P(T_{\text{new}}=0|\mathbf{x}_{\text{new}},\mathbf{w})}\right)=\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}}$$



- There are no constraints on this value: it can take any real value.

$P(T_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w}) \ll P(T_{\text{new}}=0|\mathbf{x}_{\text{new}},\mathbf{w})$ Large *negative*

$P(T_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w}) \gg P(T_{\text{new}}=0|\mathbf{x}_{\text{new}},\mathbf{w})$ Large positive

SISTA 3

From the Logit to Logistic Function

Example of a **generalized linear model**: linear model passed through a transformation to model a quantity of interest.

- Now, derive $P(T_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})$

Note: $P(T_{\text{new}}=0|\mathbf{x}_{\text{new}},\mathbf{w})=1-P(T_{\text{new}}=1|\mathbf{x}_{\text{new}},\mathbf{w})$

$$\begin{aligned}
\log\left(\frac{P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})}{P(T_{new}=0|\mathbf{x}_{new},\mathbf{w})}\right) &= \mathbf{w}^{\top}\mathbf{x}\\
\frac{P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})}{P(T_{new}=0|\mathbf{x}_{new},\mathbf{w})} &= \exp(\mathbf{w}^{\top}\mathbf{x})\\
\frac{P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})}{1-P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})} &= \exp(\mathbf{w}^{\top}\mathbf{x})\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}) &= \exp(\mathbf{w}^{\top}\mathbf{x})(1-P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}))\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}) &= \exp(\mathbf{w}^{\top}\mathbf{x})-\exp(\mathbf{w}^{\top}\mathbf{x})P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})+\exp(\mathbf{w}^{\top}\mathbf{x})P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}) &= \exp(\mathbf{w}^{\top}\mathbf{x})\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})(1+\exp(\mathbf{w}^{\top}\mathbf{x})) &= \exp(\mathbf{w}^{\top}\mathbf{x})\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}) &= \frac{\exp(\mathbf{w}^{\top}\mathbf{x})}{1+\exp(\mathbf{w}^{\top}\mathbf{x})}\\
P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}) &= \frac{1}{1+\exp(-\mathbf{w}^{\top}\mathbf{x})}
\end{aligned}$$

So the logistic function is really modeling the log-odds with a linear model!

The **Logistic** function (the inverse Logit)

4

Logistic as Likelihood

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

The Logistic (or sigmoid) function:

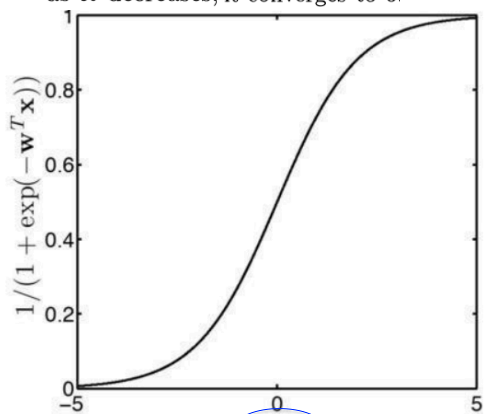
$$P(T_n = 1|\mathbf{x}_n,\mathbf{w}) = \frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x}_n)}$$

Linear component

When target is 0:

$$\begin{aligned} P(T_n = 0|\mathbf{x}_n,\mathbf{w}) &= 1 - P(T_n = 1|\mathbf{x}_n,\mathbf{w}) \\ &= 1 - \frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x}_n)} \\ &= \frac{\exp(-\mathbf{w}^\top\mathbf{x}_n)}{1+\exp(-\mathbf{w}^\top\mathbf{x}_n)}. \end{aligned}$$

As $\mathbf{w}^\top\mathbf{x}$ increases, the value converges to 1 as it decreases, it converges to 0.



Combine both into a single probability function

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

(Note! Not just fn of x)

In Sum: Problems Optimizing Logistic Regression Our analytic optimization tools (to date) fail

- **Problem 1**: cannot directly compute max likelihood (or MAP) – cannot isolate **w**

$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} \left(\frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x}_n)}\right)^{t_n} \left(\frac{\exp(-\mathbf{w}^\top\mathbf{x}_n)}{1+\exp(-\mathbf{w}^\top\mathbf{x}_n)}\right)^{1-t_n}$$

- **Problem 2**: cannot compute full Bayes because cannot integrate Bayes denominator (marginal likelihood)

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

SISTA 6

Numerator of Bayes Theorem
$$g(\mathbf{w};\mathbf{X},\mathbf{t},\sigma^2) = \underbrace{p(\mathbf{t}|\mathbf{X},\mathbf{w})}_{\text{likelihood}}\underbrace{p(\mathbf{w}|\sigma^2)}_{\text{prior}}$$

marginal Likelihood (denominator)
$$Z = p(\mathbf{t}|\mathbf{X},\sigma^2) = \int p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\sigma^2)d\mathbf{w}$$

The Posterior
$$p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2) = Z^{-1}g(\mathbf{w};\mathbf{X},\mathbf{t},\sigma^2)$$

Our Options

1. Find the single value of **w** that corresponds to the highest value of the likelihood or posterior. As *g* is proportional to the posterior, a maximum of *g* will also correspond to a maximum of the posterior. Z^{-1} is not a function of **w**
2. Approximate $p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)$ with some other density that we can compute analytically.
3. Sample directly from the posterior $p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)$, knowing only $g(\mathbf{w};\mathbf{X},\mathbf{t},\sigma^2)$

SISTA 7

Method 1: point estimate

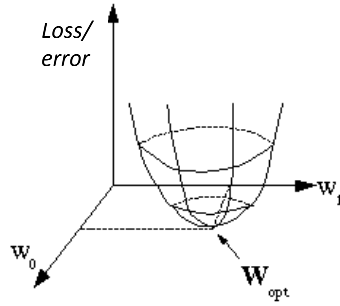
- We can find **w** that maximizes the likelihood (MLE) or the posterior (MAP).
- **Consider MAP**: while we cannot derive a direct analytic posterior density, we can compute something proportional to it:

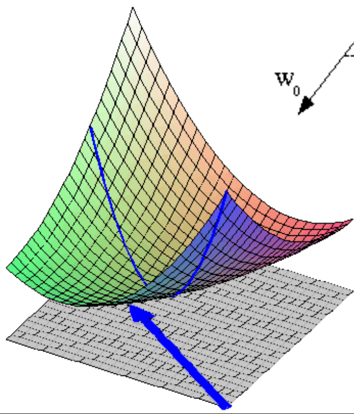
$$g(\mathbf{w};\mathbf{X},\mathbf{t},\sigma^2) = p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\sigma^2)$$

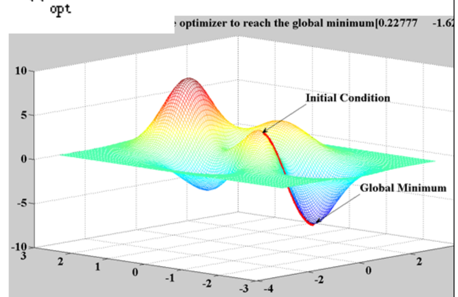
- We will find the value of **w** that maximizes *g*
- This will correspond to the value at the maximum of the posterior.
- This will be the most likely value $\hat{\mathbf{w}}$ under the posterior.

SISTA 8

Gradient Methods: Descent/Ascent





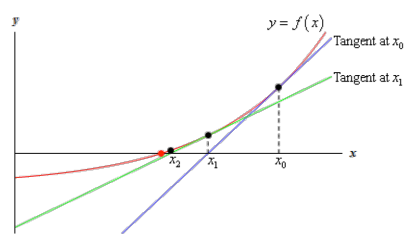


The parameter update rule (Widrow-Hoff)
(applied to linear least squares)

- The ***batch*** update version

$$\begin{aligned} \mathbf{w} &:= \mathbf{w} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \\ &:= \mathbf{w} - \alpha \left(\mathbf{X}^\top \mathbf{X} \mathbf{w} - \mathbf{X}^\top \mathbf{t} \right) \\ &:= \mathbf{w} - \alpha \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \mathbf{x}_n \right) \mathbf{x}_n \end{aligned}$$

Properties:
(1) Step is in the direction of the gradient's steepest descent (negative of tangent slope)
(2) Magnitude of the update is proportional to the error term and scaled by α



The "algorithm"

repeat
$\quad \mathbf{w} := \mathbf{w} - \alpha \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \mathbf{x}_n \right) \mathbf{x}_n$
until *convergence*;

Version with design matrix computations (equivalent)

repeat
$\quad \mathbf{w} := \mathbf{w} - \alpha \left(\mathbf{X}^\top \mathbf{X} \mathbf{w} - \mathbf{X}^\top \mathbf{t} \right)$
until *convergence*;

SISTA 10

The Newton-Raphson Method

Isaac Newton 1642–1727 · Joseph Raphson 1648–1715

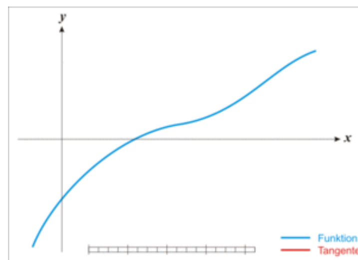
- **Newton's Method**: finding points where functions are equal to zero (e.g., ***roots*** of a polynomial)
- Given a current estimate of the zero point, x_n, find derivative at that point to get the slope, find intersection of sloping line, then move in that direction:

f at current pt · *f* at goal *root*

$$f'(x_n) = \frac{\Delta y}{\Delta x} = \frac{f(x_n) - 0}{x_n - x_{n+1}}.$$

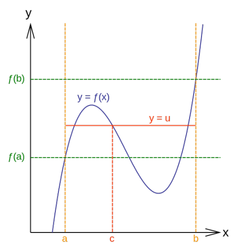
Location in x we'd like to know

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$



Why does this work?
The **intermediate value theorem**!!

If $f : [a,b] \rightarrow \mathbf{R}$ is continuous, u is real and $f(a) > u > f(b)$, **then** $\exists\, c \in (a,b),\ f(c) = u$



SISTA 11

The Newton-Raphson Method

- But this method can do more!
- Extend to find minima and maxima of a function.
- These are simply points where the first derivative itself passes through zero.
- So… replace *f* with *f'* and *f'* with *f''*

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

SISTA 12

Using Newton-Raphson for MAP logistic regression

$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N}\left(\frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x}_n)}\right)^{t_n}\left(\frac{\exp(-\mathbf{w}^\top \mathbf{x}_n)}{1+\exp(-\mathbf{w}^\top \mathbf{x}_n)}\right)^{1-t_n} \qquad p(\mathbf{w}|\sigma^2) = \mathcal{N}(0, \sigma^2\mathbf{I})$$

$$\log g(\mathbf{w};\mathbf{X},\mathbf{t}) = \log p(\mathbf{t}|\mathbf{X},\mathbf{w}) + \log p(\mathbf{w}|\sigma^2)$$

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

$$\mathbf{w}' = \mathbf{w} - \left(\frac{\partial^2 \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}}\right)^{-1}\frac{\partial \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}}$$

Point is a **global maximum** if Hessian is negative definite
(as was the case with the linear additive Gaussian noise max likelihood)

13

$$\mathbf{w}' = \mathbf{w} - \left(\frac{\partial^2 \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}}\right)^{-1}\frac{\partial \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}}$$

Take the Derivatives…

$$P_n = P(T_n = 1|\mathbf{w},\mathbf{x}_n) = \frac{1}{1+\exp(\mathbf{w}^\top \mathbf{x}_n)}$$

Recall the chain rule:

$$\frac{\partial f(g(\mathbf{w}))}{\partial \mathbf{w}} = \frac{\partial f(g(\mathbf{w}))}{\partial g(\mathbf{w})}\frac{\partial g(\mathbf{w})}{\partial \mathbf{w}}$$

$$\log g(\mathbf{w};\mathbf{X},\mathbf{t}) = \log p(\mathbf{t}|\mathbf{X},\mathbf{w}) + \log p(\mathbf{w}|\sigma^2)$$

$$\begin{aligned}\log g(\mathbf{w};\mathbf{X},\mathbf{t}) &= \sum_{n=1}^{N}\log P(T_n = t_n|\mathbf{x}_n,\mathbf{w}) + \log p(\mathbf{w}|\sigma^2)\\
&= \sum_{n=1}^{N}\log\left(\frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}\right)^{t_n}\left(\frac{\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}\right)^{1-t_n}\\
&\quad + \log p(\mathbf{w}|\sigma^2)\end{aligned}$$

$$\begin{aligned}&= \log p(\mathbf{w}|\sigma^2) + \sum_{n=1}^{N}\log P_n^{t_n} + \log(1-P_n)^{1-t_n}\\
&= -\frac{D}{2}\log 2\pi - D\log\sigma - \frac{1}{2\sigma^2}\mathbf{w}^\mathsf{T}\mathbf{w} \\
&\quad + \sum_{n=1}^{N} t_n \log P_n + (1-t_n)\log(1-P_n),\end{aligned}$$

Log of (zero mean) Gaussian prior

14

$$\mathbf{w}' = \mathbf{w} - \left(\frac{\partial^2 \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}}\right)^{-1} \frac{\partial \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}}$$

Take the Derivatives…

$$P_n = P(T_n = 1|\mathbf{w},\mathbf{x}_n) = \frac{1}{1+\exp(\mathbf{w}^\top\mathbf{x}_n)}$$

Recall the chain rule:

$$\frac{\partial f(g(\mathbf{w}))}{\partial\mathbf{w}} = \frac{\partial f(g(\mathbf{w}))}{\partial g(\mathbf{w})}\frac{\partial g(\mathbf{w})}{\partial\mathbf{w}}$$

$$\log g(\mathbf{w};\mathbf{X},\mathbf{t}) = -\frac{D}{2}\log 2\pi - D\log\sigma - \frac{1}{2\sigma^2}\mathbf{w}^\mathsf{T}\mathbf{w} + \sum_{n=1}^{N} t_n \log P_n + (1-t_n)\log(1-P_n)$$

$$\begin{aligned}\frac{\partial \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}} &= -\frac{1}{\sigma^2}\mathbf{w} + \sum_{n=1}^{N}\left(\frac{t_n}{P_n}\frac{\partial P_n}{\partial\mathbf{w}} + \frac{1-t_n}{1-P_n}\frac{\partial(1-P_n)}{\partial\mathbf{w}}\right)\\ &= -\frac{1}{\sigma^2}\mathbf{w} + \sum_{n=1}^{N}\left(\frac{t_n}{P_n}\frac{\partial P_n}{\partial\mathbf{w}} - \frac{1-t_n}{1-P_n}\frac{\partial P_n}{\partial\mathbf{w}}\right),\end{aligned}$$

Apply chain rule again…

$$\begin{aligned}\frac{\partial(1-P_n)}{\partial\mathbf{w}} &= \frac{\partial(1-P_n)}{\partial P_n}\frac{\partial P_n}{\partial\mathbf{w}}\\ &= -\frac{\partial P_n}{\partial\mathbf{w}}.\end{aligned}$$

$$\begin{aligned}\frac{\partial P_n}{\partial\mathbf{w}} &= \frac{\partial\,(1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n))^{-1}}{\partial\,(1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n))}\frac{\partial\,(1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n))}{\partial\mathbf{w}}\\ &= -\frac{1}{(1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n))^2}\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)(-\mathbf{x}_n)\\ &= \frac{\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}{(1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n))^2}\mathbf{x}_n\\ &= \frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}\frac{\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x})}\mathbf{x}_n\\ &= P_n(1-P_n)\mathbf{x}_n.\end{aligned}$$

Finally, plug $\frac{\partial P_n}{\partial\mathbf{w}}$ back in:

$$\begin{aligned}\frac{\partial \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}} &= -\frac{1}{\sigma^2}\mathbf{w} + \sum_{n=1}^{N}(\mathbf{x}_n t_n(1-P_n) - \mathbf{x}_n(1-t_n)P_n)\\ &= -\frac{1}{\sigma^2}\mathbf{w} + \sum_{n=1}^{N}\mathbf{x}_n(t_n - t_nP_n - P_n + t_nP_n)\\ &= -\frac{1}{\sigma^2}\mathbf{w} + \sum_{n=1}^{N}\mathbf{x}_n(t_n - P_n).\end{aligned}$$

$$\mathbf{w}' = \mathbf{w} - \left(\frac{\partial^2 \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}}\right)^{-1} \frac{\partial \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}}$$

Take the Derivatives…

$$P_n = P(T_n = 1|\mathbf{w},\mathbf{x}_n) = \frac{1}{1+\exp(\mathbf{w}^\top\mathbf{x}_n)}$$

Recall the chain rule:

$$\frac{\partial f(g(\mathbf{w}))}{\partial\mathbf{w}} = \frac{\partial f(g(\mathbf{w}))}{\partial g(\mathbf{w})}\frac{\partial g(\mathbf{w})}{\partial\mathbf{w}}$$

This is the first derivative!

$$\frac{\partial \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}} = -\frac{1}{\sigma^2}\mathbf{w} + \sum_{n=1}^{N}\mathbf{x}_n(t_n - P_n)$$

Now, for the Hessian…:

$$\begin{aligned}\frac{\partial^2 \log g(\mathbf{w};\mathbf{X},\mathbf{t})}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}} &= -\frac{1}{\sigma^2}\mathbf{I} - \sum_{n=1}^{N}\mathbf{x}_n\frac{\partial P_n}{\partial\mathbf{w}^\mathsf{T}}\\ &= -\frac{1}{\sigma^2}\mathbf{I} - \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\mathsf{T}P_n(1-P_n)\end{aligned}$$

$$\frac{\partial P_n}{\partial\mathbf{w}} = P_n(1-P_n)\mathbf{x}_n$$

$$\frac{\partial P_n}{\partial\mathbf{w}^\mathsf{T}} = \left(\frac{\partial P_n}{\partial\mathbf{w}}\right)^\mathsf{T}$$

$0 \le P_n \le 1$ So, Hessian is negative definite!

There can only be one optimum, must be a maximum.
Whatever value of **w** the Newton-Raphson procedure converges to must correspond to the highest value of the posterior density.
This is a choice of our particular prior and likelihood
Changing either may result in harder posterior density to optimize

16

The final Newton-Raphson update rule
(for Bayesian logistic regression)

This just means "evaluate with": plug in $\mathbf{w}_t$ for all occurrences of $\mathbf{w}$ after taking the derivative

$$
\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t - \frac{f'(\mathbf{w}_t)}{f''(\mathbf{w}_t)} \\
&= \mathbf{w}_t - \left(\frac{\partial^2 \log g(\mathbf{w}; \mathbf{Xt})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \bigg|_{\mathbf{w}_t} \right)^{-1} \left(\frac{\partial \log g(\mathbf{w}; \mathbf{Xt})}{\partial \mathbf{w}} \bigg|_{\mathbf{w}_t} \right) \\
&= \mathbf{w}_t - \left(-\frac{1}{\sigma^2}\mathbf{I} - \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top P_n(1-P_n) \right)^{-1} \left(-\frac{1}{\sigma^2}\mathbf{w}_t + \sum_{n=1}^{N} \mathbf{x}_n (t_n - P_n) \right)
\end{aligned}
$$

where

$$P_n = P(T_n = 1 | \mathbf{w}, \mathbf{x}_n) = \frac{1}{1 + \exp(\mathbf{w}^\top \mathbf{x}_n)}$$

SISTA 17

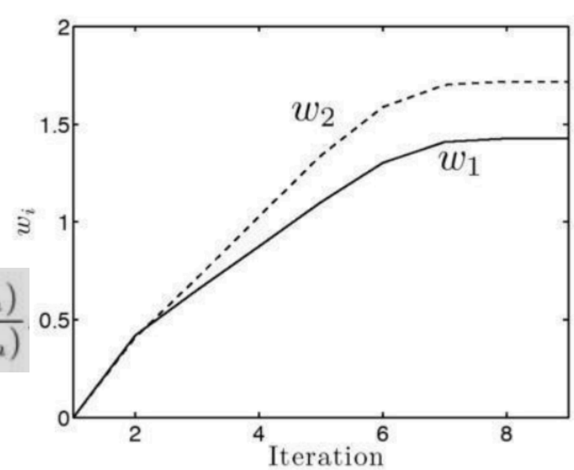
Estimating w

Starting from:

$$\mathbf{w} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\sigma^2 = 10$$

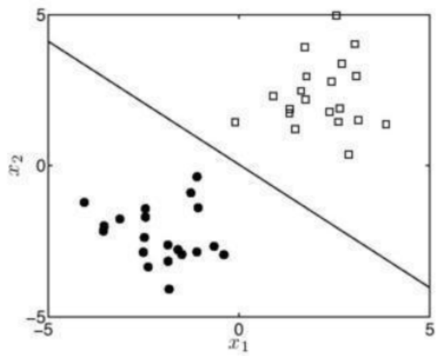
$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

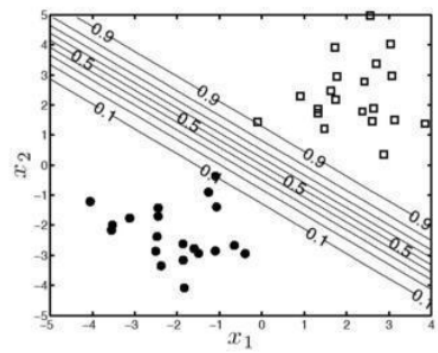


SISTA 18

Using w to compute prob of response

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

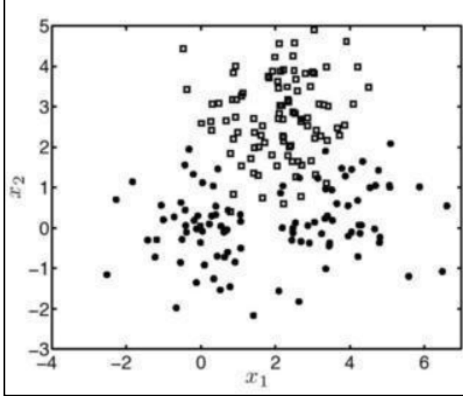


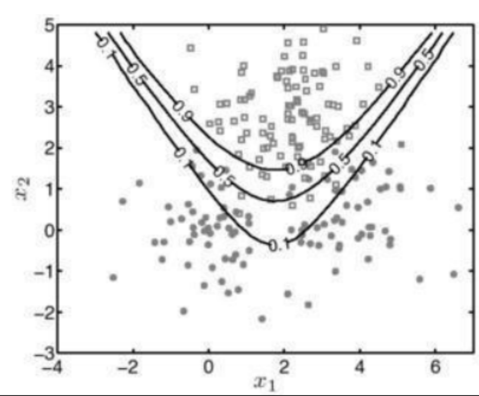


Nonlinear Decision Functions

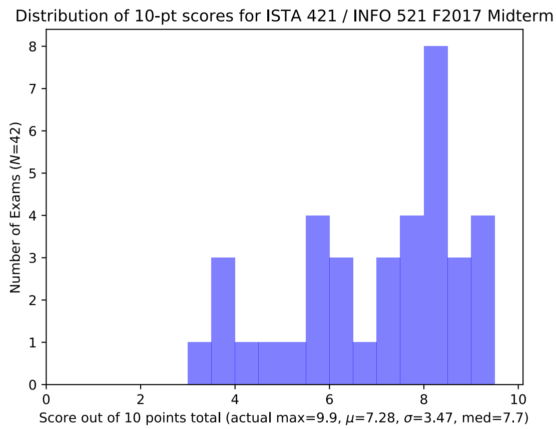
$$\log\left(\frac{P(T_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{w})}{P(T_{\text{new}} = 0 | \mathbf{x}_{\text{new}}, \mathbf{w})}\right) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2$$

Find $\widehat{\mathbf{w}}$ by MAP





Midterm 10pt score distribtuion



21