Logistic Regression

Aarti Singh

Machine Learning 10-315
Sept 15, 2021

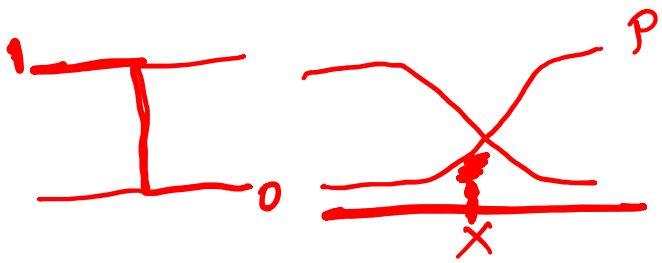
Announcements

- QnA1 grades out

- Recitation Friday Sept 17
  - Optimization

- Anonymous feedback
  - Voice
  - Recitation
  - More math/practice problems
  - More relevant HW
  - Lecture notes

Practice problems



- True/False: Bayes optimal classifier has 0 error rate. (QnA1)
- MLE for multinomial
  - Prove likelihood is concave
- MAP for Bernoulli – show that Beta is a conjugate prior i.e. if prior is Beta(β_H,β_T), posterior is Beta(β_H+α_H,β_T+α_T)
  - Find mode of Beta distribution
- [Advanced] MAP for multinomial using Dirichlet conjugate prior
  - Find mode of Dirichlet distribution
- MLE for Gaussian mean in d-dimensions - Prove likelihood is concave
- MLE for Gaussian variance in 1-dimension - Prove likelihood is concave
- MLE for Gaussian covariance in d-dimensions - Prove likelihood is concave
- MAP for Gaussian mean (1 and d-dimensions) – show that Gaussian is a conjugate prior for mean μ i.e. if prior for mean is Gaussian(η,λ^2), posterior is Gaussian - Find parameters of posterior Gaussian, mode of posterior
- [Advanced] MAP for Gaussian variance (1 and d-dimensions) – Find mode of posterior

Practice problems

- MLE for uniform distribution - Prove likelihood is concave
- MLE for exponential distribution - Prove likelihood is concave
- When is MLE same as MAP?
- Implement Naive Bayes algorithm on real data (HW2)
- Comparison between different methods (assumptions, number of parameters, decision boundary)
  - Bayes vs Naïve Bayes
  - Gaussian Bayes vs Gaussian Naïve Bayes

Discriminative Classifiers

Bayes Classifier:

$$f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$$
$$= \arg\max_{Y=y} P(X = x|Y = y)P(Y = y)$$

Generative

Why not learn P(Y|X) directly? Or better yet, why not learn the decision boundary directly?

- Assume some functional form for P(Y|X) or for the decision boundary
- Estimate parameters of functional form directly from training data

Today we will see one such classifier – **Logistic Regression**

Logistic Regression

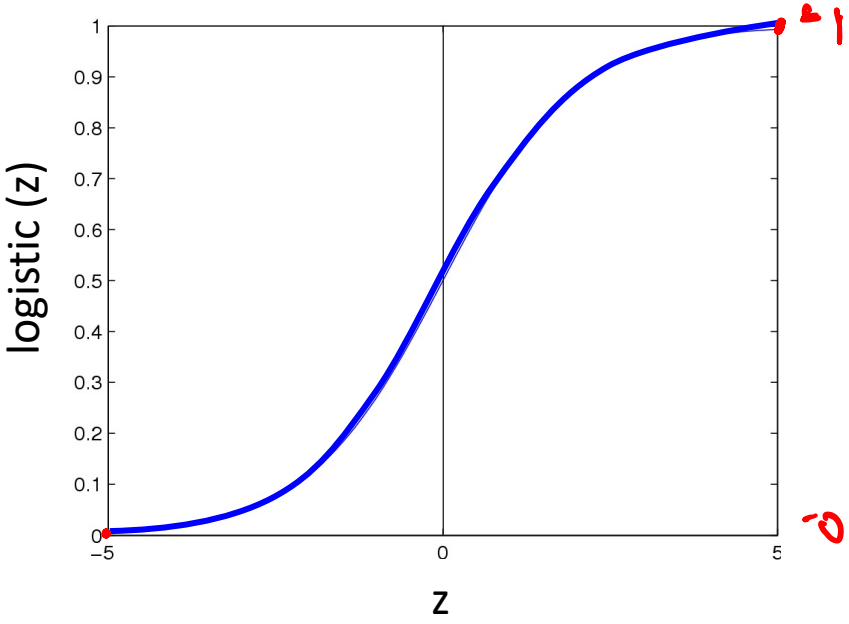
Not really regression

Assumes the following functional form for P(Y|X):

$$P(Y=1|X) = \frac{1}{1+\exp(-w_0 - \sum_i w_i X_i)}$$

Logistic function applied to a linear function of the data

Logistic function (or Sigmoid), σ(z) = $\frac{1}{1+exp(-z)}$

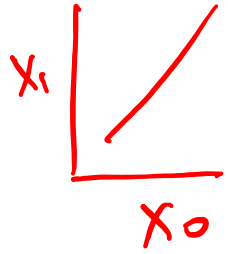


Features can be discrete or continuous!

Logistic Regression is a Linear Classifier!

Assumes the following functional form for P(Y|X):

$$P(Y=1|X) = \frac{1}{1+\exp(-w_0 - \sum_i w_i X_i)}$$



$$\Rightarrow P(Y=0|X) = 1 - \frac{1}{1+e^{-z}} = \frac{1+e^{-z}-1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}} = \frac{1}{1+e^{z}}$$

Decision boundary:

$$\Rightarrow \frac{P(Y=1|X)}{P(Y=0|X)} = \exp(w_0 + \sum_i w_i X_i) \gtrless 1 \quad (e^z)$$

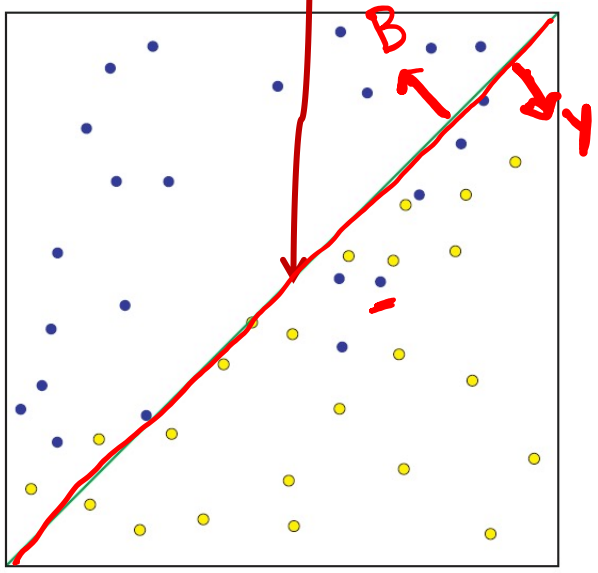
(1 above, 0 below the $\gtrless$)

$$\Rightarrow w_0 + \sum_i w_i X_i \gtrless 0$$

(1 above, 0 below the $\gtrless$)

(Linear Decision Boundary)

$$w_0 + \sum_i w_i X_i = 0$$



Training Logistic Regression

How to learn the parameters $w_0, w_1, \ldots w_d$? (d features)

Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

Maximum Likelihood Estimates

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{n} P(X^{(j)}, Y^{(j)} \mid \mathbf{w})$$

$$\propto P(Y|X)\,P(X) \propto P(X|Y)\,P(Y)$$

But there is a problem …

Don't have a model for P(X) or P(X|Y) – only for P(Y|X)

Training Logistic Regression

How to learn the parameters $w_0, w_1, \ldots w_d$? (d features)

Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$

Maximum (Conditional) Likelihood Estimates

$$\hat{\mathbf{w}}_{MCLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{n} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})$$

Discriminative philosophy – Don't waste effort learning P(X), focus on P(Y|X) – that's all that matters for classification!

Expressing Conditional log Likelihood

$$P(Y=1|X,w) = \frac{1}{1+\exp(-w_0 - \sum_i w_i X_i)} = \frac{e^{w_0+\sum_i w_i X_i}}{1+e^{w_0+\sum_i w_i X_i}}$$

$$P(Y=0|X,w) = \frac{1}{1+\exp(w_0 + \sum_i w_i X_i)}$$

$$l(\mathbf{w}) \equiv \ln \prod_j P(y^j|\mathbf{x}^j,\mathbf{w})$$

$$P(Y=y|X,w) = \frac{e^{y(\sum_i w_i x_i + w_0)}}{1+e^{(\sum_i w_i x_i + w_0)}}$$

$$= \sum_j \ln P(y^j|x^j,w) \quad *$$

$$= \sum_j \ln \frac{e^{y^j(w_0+\sum_i w_i x_i^j)}}{1+e^{(\ )}} = \sum_j y^j(w_0+\sum_i w_i x_i^j) - \ln(1+e^{(\ )}) \quad \#$$

$* \ln(ab) = \ln a + \ln b$ $\quad \# \ln \frac{a}{b} = \ln a - \ln b$

Expressing Conditional log Likelihood

$$P(Y = 1|X, w) = \frac{1}{1 + \exp(-w_0 - \sum_i w_i X_i)}$$

$$P(Y = 0|X, w) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

$$\begin{aligned} l(\mathbf{w}) &\equiv \ln \prod_j P(y^j|\mathbf{x}^j, \mathbf{w}) \\ &= \sum_j \left[y^j (w_0 + \sum_i^d w_i x_i^j) - \ln(1 + exp(w_0 + \sum_i^d w_i x_i^j)) \right] \end{aligned}$$

Good news: $l(\mathbf{w})$ is concave function of $\mathbf{w}$! QnA2

Expressing Conditional log Likelihood

$$P(Y=1|X,w) = \frac{1}{1+\exp(-w_0 - \sum_i w_i X_i)}$$

$$P(Y=0|X,w) = \frac{1}{1+\exp(w_0 + \sum_i w_i X_i)}$$

$$\begin{aligned} l(\mathbf{w}) &\equiv \ln \prod_j P(y^j|\mathbf{x}^j,\mathbf{w}) \\ &= \sum_j \left[y^j (w_0 + \sum_i^d w_i x_i^j) - \ln(1 + exp(w_0 + \sum_i^d w_i x_i^j)) \right] \end{aligned}$$

Good news: *l*(**w**) is concave function of **w !** QnA2

Bad news: no closed-form solution to maximize *l*(**w**)

Good news: can use iterative optimization methods (gradient ascent)

That's M(C)LE. How about M(C)AP?

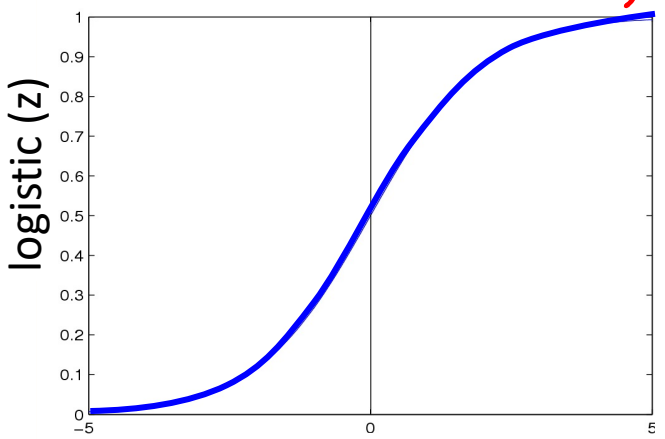
$$p(\mathbf{w} \mid Y, \mathbf{X}) \propto P(Y \mid \mathbf{X}, \mathbf{w}) p(\mathbf{w})$$

- Define priors on **w**
  - Common assumption: Normal distribution, zero mean, identity covariance
  - "Pushes" parameters towards zero

$$p(\mathbf{w}) = \prod_i \frac{1}{\kappa\sqrt{2\pi}} e^{\frac{-w_i^2}{2\kappa^2}}$$

Zero-mean Gaussian prior

Logistic function (or Sigmoid), σ(z) = : $\frac{1}{1 + exp(-z)}$



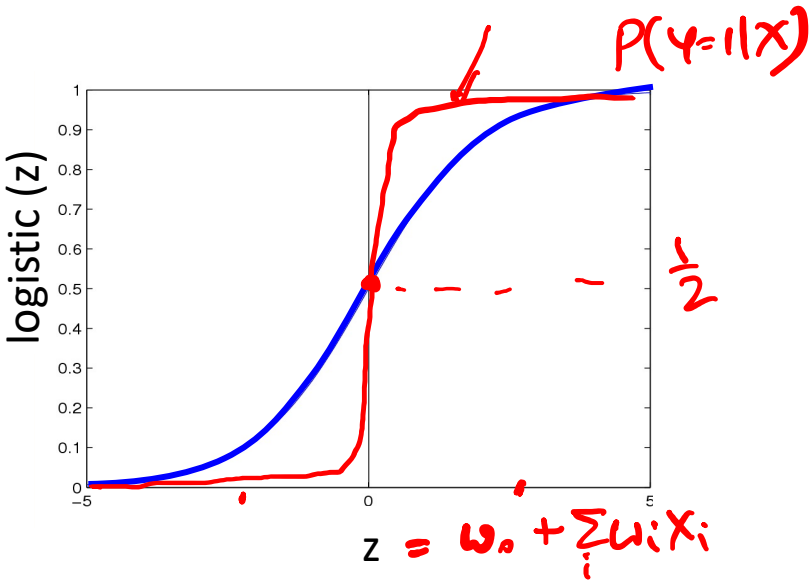
➢ What happens if we scale z by a large constant?

Logistic function (or Sigmoid), $\sigma(z) = :$

$$\frac{1}{1 + exp(-z)}$$

$p(Y=1|X)$

$z = w_0 + \sum_i w_i x_i \geq 0$



- Poll: What happens if we scale z (equivalently weights w) by a large constant?

A) The logistic classifier decision boundary shifts towards class 1
B) The logistic classifier decision boundary remains same ✓
C) The logistic classifier tries to separate the data perfectly ✓
D) The logistic classifier allows more mixing of labels on each side of decision boundary

That's M(C)LE. How about M(C)AP?

$$p(\mathbf{w} \mid Y, \mathbf{X}) \propto P(Y \mid \mathbf{X}, \mathbf{w}) p(\mathbf{w})$$

$$p(\mathbf{w}) = \prod_i \frac{1}{\kappa\sqrt{2\pi}} e^{\frac{-w_i^2}{2\kappa^2}}$$

Zero-mean Gaussian prior

- M(C)AP estimate

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ln \left[p(\mathbf{w}) \prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w}) \right]$$

$$= \ln p(\mathbf{w}) + \ln \prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w})$$

log prior + log conditional likelihood

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \sum_{j=1}^{n} \ln P(y^j \mid \mathbf{x}^j, \mathbf{w}) - \sum_{i=1}^{d} \frac{w_i^2}{2\kappa^2}$$

$$\ln \prod_i e^{-w_i^2/2\kappa^2} = \sum_i -\frac{w_i^2}{2\kappa^2}$$

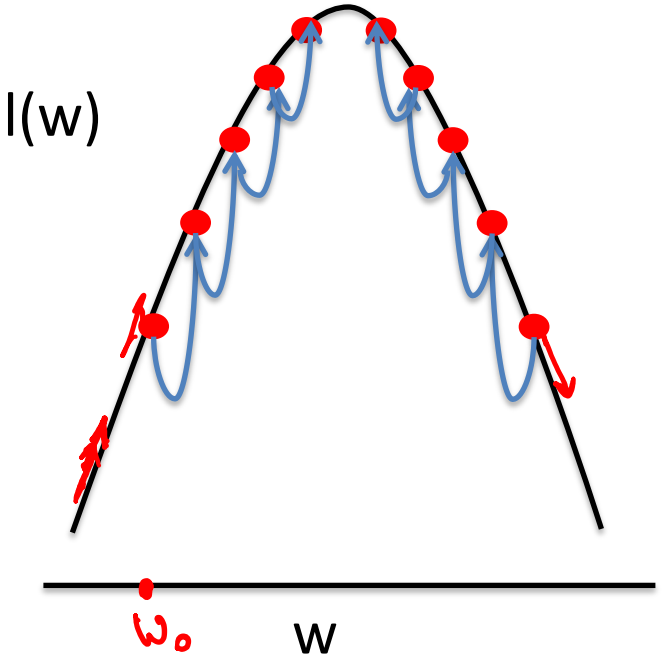
Still concave objective!

Penalizes large weights

Iteratively optimizing concave function

- Conditional likelihood for Logistic Regression is concave
- Maximum of a concave function can be reached by

Gradient Ascent Algorithm



Initialize: Pick **w** at random

Gradient:

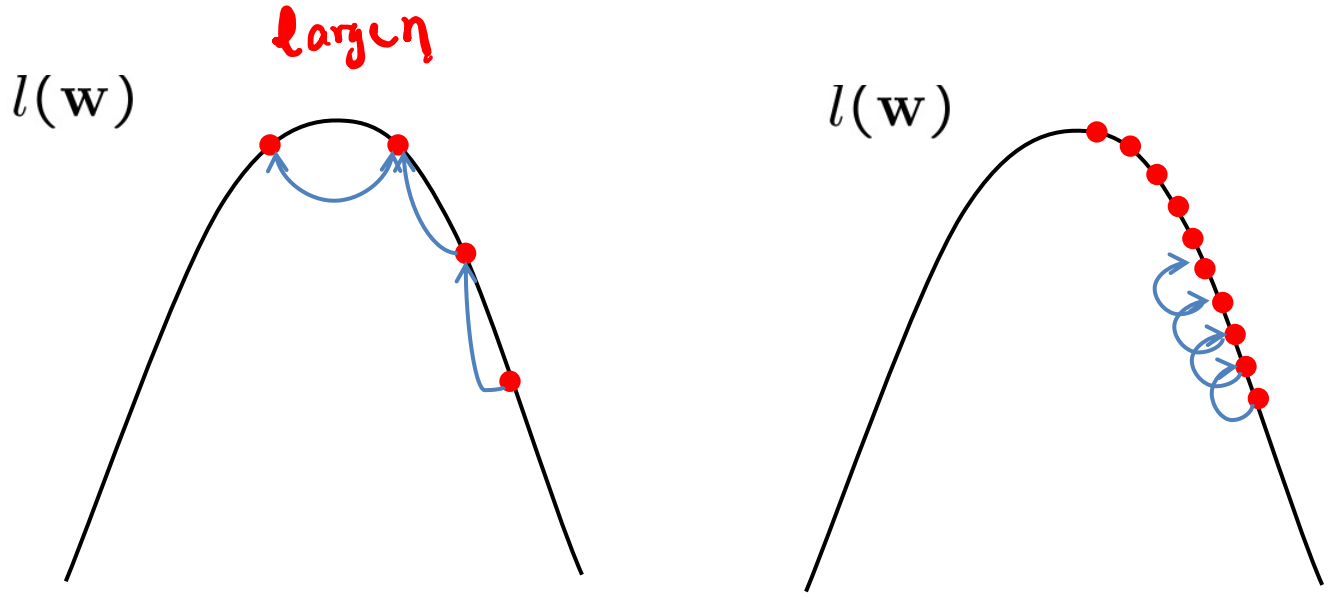
$$\nabla_{\mathbf{w}} l(\mathbf{w}) = [\frac{\partial l(\mathbf{w})}{\partial w_0}, \ldots, \frac{\partial l(\mathbf{w})}{\partial w_d}]'$$

Update rule: Learning rate, $\eta > 0$

$$\Delta \mathbf{w} = \eta \nabla_{\mathbf{w}} l(\mathbf{w})$$

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \left. \frac{\partial l(\mathbf{w})}{\partial w_i} \right|_t$$

Effect of step-size η



Large η => Fast convergence but larger residual error
Also possible oscillations

Small η => Slow convergence but small residual error

Gradient Ascent for M(C)LE

Gradient ascent rule for w_0:

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \left. \frac{\partial l(\mathbf{w})}{\partial w_0} \right|_t$$

$$l(\mathbf{w}) = \sum_j \left[y^j (w_0 + \sum_i^d w_i x_i^j) - \ln(1 + exp(w_0 + \sum_i^d w_i x_i^j)) \right]$$

$$\frac{\partial}{\partial w_0} = \sum_j \left[y^j - \frac{e^{(\,)}}{1+e^{(\,)}} \right]$$

$$(\,) = w_0 + \sum_i w_i x_i^j$$

$$P(Y=1 \mid X, \mathbf{w}^t)$$

$$\frac{\partial}{\partial w_i} = \sum_j x_i^j \left[y^j - \frac{e^{(\,)}}{1+e^{(\,)}} \right]$$

$$P(Y=1 \mid X, \mathbf{w}) = \frac{1}{1+e^{-(\,)}} = \frac{e^{(\,)}}{1+e^{(\,)}}$$

Gradient Ascent for M(C)LE

Gradient ascent rule for w_0:

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \left. \frac{\partial l(\mathbf{w})}{\partial w_0} \right|_t$$

$$l(\mathbf{w}) = \sum_j \left[y^j (w_0 + \sum_i^d w_i x_i^j) - \ln(1 + exp(w_0 + \sum_i^d w_i x_i^j)) \right]$$

$$\frac{\partial l(\mathbf{w})}{\partial w_0} = \sum_j \left[y^j - \underbrace{\frac{1}{1 + exp(w_0 + \sum_i^d w_i x_i^j)} \cdot exp(w_0 + \sum_i^d w_i x_i^j)}_{} \right]$$

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \sum_j [y^j - \hat{P}(Y^j = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})]$$

Gradient Ascent for M(C)LE Logistic Regression

Gradient ascent algorithm: iterate until change < ε

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \sum_j [y^j - \hat{P}(Y^j = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})]$$

For i=1,…,d,

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \sum_j x_i^j [y^j - \hat{P}(Y^j = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})]$$

repeat

Predict what current weight thinks label Y should be

- Gradient ascent is simplest of optimization approaches
  - e.g. Stochastic gradient ascent, Momentum methods, Newton method, Conjugate gradient ascent, IRLS (see Bishop 4.3.3)

M(C)AP – Gradient

- Gradient

$$p(\mathbf{w}) = \prod_i \frac{1}{\kappa\sqrt{2\pi}} \; e^{\frac{-w_i^2}{2\kappa^2}}$$

Zero-mean Gaussian prior

$$\frac{\partial}{\partial w_i} \ln \left[p(\mathbf{w}) \prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w}) \right]$$

$$\ln p(w) \propto \sum_i -\frac{w_i^2}{2\kappa^2}$$

$$\frac{\partial}{\partial w_i} \ln p(\mathbf{w}) + \frac{\partial}{\partial w_i} \ln \left[\prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w}) \right]$$

Same as before

$$\frac{\partial}{\partial w_i} \ln p(w) \propto -\sum_i \frac{w_i}{\kappa^2}$$

$$\propto \frac{-w_i}{\kappa^2}$$

Extra term Penalizes large weights

Penalization = Regularization

M(C)LE vs. M(C)AP

- Maximum conditional likelihood estimate

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ln \left[\prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w}) \right]$$

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \sum_j x_i^j [y^j - P(Y = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})]$$

- Maximum conditional a posteriori estimate

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ln \left[p(\mathbf{w}) \prod_{j=1}^{n} P(y^j \mid \mathbf{x}^j, \mathbf{w}) \right]$$

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \left\{ -\frac{1}{\kappa^2} w_i^{(t)} + \sum_j x_i^j [y^j - P(Y = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})] \right\}$$

Logistic Regression for more than 2 classes

- Logistic regression in more general case, where $Y \in \{y_1, \ldots, y_K\}$

for $k<K$

$$P(Y = y_k|X) = \frac{\exp(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i)}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji} X_i)}$$

for $k=K$ (normalization, so no weights for this class)

$$P(Y = y_K|X) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji} X_i)}$$

Predict $f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$

Is the decision boundary still linear?