6.867 Machine learning and neural networks

Tommi Jaakkola
MIT AI Lab
tommi@ai.mit.edu

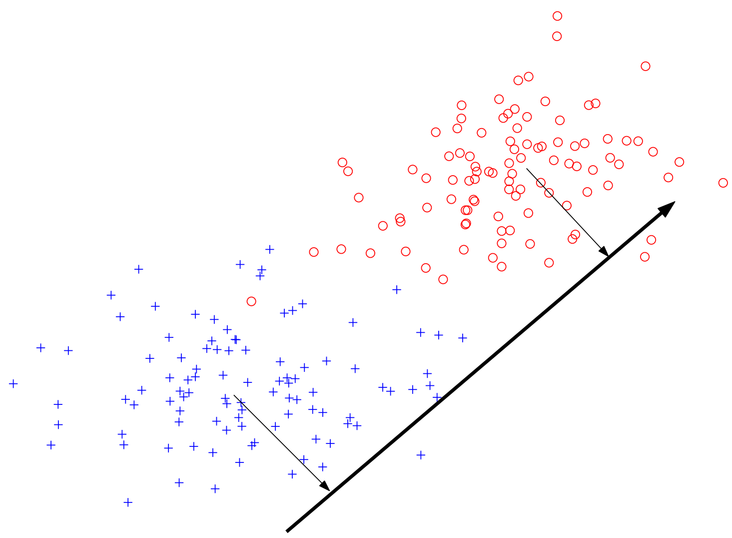
Lecture 4: classification

Topics

- Classification
  - Fisher linear discriminant analysis
  - Generative probabilistic classifiers
  - Discriminative classifiers, additive models

Beyond regression: Fisher linear discriminant analysis

- Assume two sets of examples (classes 1 and 0) with means μ_1, μ_0 and covariances Σ_1, Σ_0 (not necessarily normal).

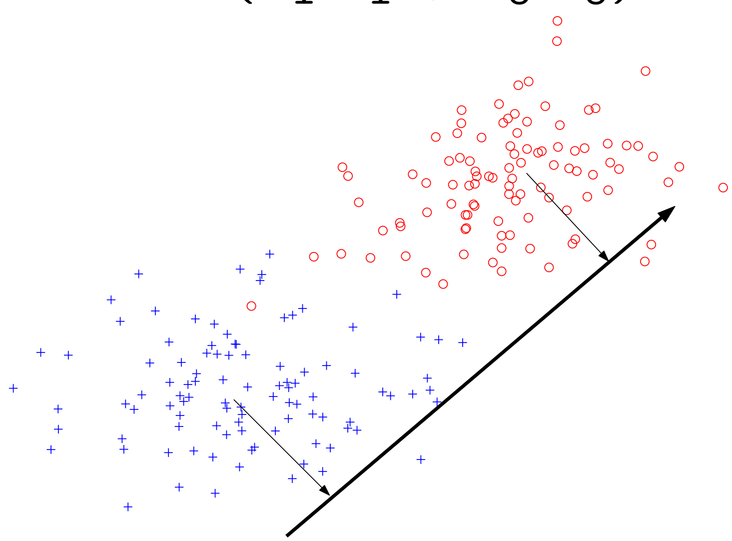


- We try to find a direction $\mathbf{w} = [w_1, \ldots, w_d]^T$ in the input space such that projecting the sets along this dimension makes them "well-separated".

Fisher linear discriminant analysis cont'd

- More mathematically: we find a direction $\mathbf{w}$ (linear projection) that maximizes

$$J_{Fisher}(\mathbf{w}) = \frac{(\text{Separation of projected means})^2}{\text{Sum of within population variances}}$$
$$= \frac{(\mathbf{w}^T\mu_1 - \mathbf{w}^T\mu_0)^2}{\mathbf{w}^T(n_1\Sigma_1 + n_0\Sigma_0)\mathbf{w}}$$



- The solution is $\mathbf{w} = (n_1\Sigma_1 + n_0\Sigma_0)^{-1}(\mu_1 - \mu_0)$
  - *optimal* for two normal (Gaussian) populations with equal co-variances ($\Sigma_1 = \Sigma_0$)

Background: projected examples

- The mean and the covariance of the examples in class 1 are

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n_1} \sum_{i \in \text{class } 1} \mathbf{x}_i \\ \hat{\Sigma}_1 &= \frac{1}{n_1} \sum_{i \in \text{class } 1} (\mathbf{x}_i - \hat{\mu}_1)(\mathbf{x}_i - \hat{\mu}_1)^T\end{aligned}$$

  and similarly for $\hat{\mu}_0$ and $\hat{\Sigma}_0$. Here n_i for $i = 0, 1$ denote the number of examples in each class.

- When we project each example $\mathbf{x}_i$ along $\mathbf{w}$, we get two one dimensional sets of examples (projected examples denoted by $z_i(\mathbf{w})$). We can compute the means $\hat{m}$ and variances $\hat{\sigma}^2$ of these new examples within each class:

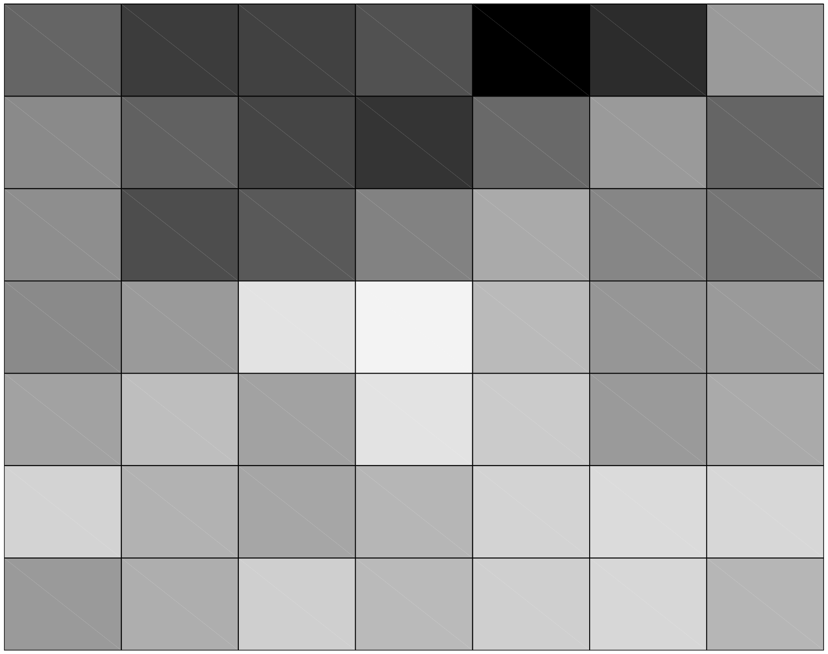
$$z_i(\mathbf{w}) = \mathbf{w}^T \mathbf{x}_i, \quad \hat{m}_1(\mathbf{w}) = \mathbf{w}^T \hat{\mu}_1, \quad \hat{\sigma}_1^2(\mathbf{w}) = \mathbf{w}^T \hat{\Sigma}_1 \mathbf{w}$$

- In Fisher discriminant analysis, we maximize

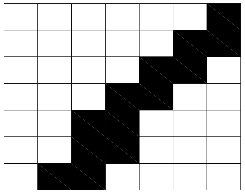
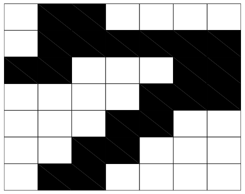
$$J_{Fisher}(\mathbf{w}) = \frac{(\hat{m}_1(\mathbf{w}) - \hat{m}_0(\mathbf{w}))^2}{n_1 \hat{\sigma}_1^2(\mathbf{w}) + n_0 \hat{\sigma}_0^2(\mathbf{w})} = \frac{(\mathbf{w}^T \hat{\mu}_1 - \mathbf{w}^T \hat{\mu}_0)^2}{\mathbf{w}^T (n_1 \hat{\Sigma}_1 + n_0 \hat{\Sigma}_0) \mathbf{w}}$$

Fisher linear discriminant analysis: example

- Binary digits "1" versus "7"



This is approximately the matrix difference "1" - "7"

$\approx$  - 

Generative and discriminative classification

- We can try to make classification decisions in two ways

  1. Generative ($\approx P(\mathbf{x}|y)$)
     - Build a model over the input examples in each class and classify based on how well the resulting class conditional models explain any new input example
  2. Discriminative ($\approx P(y|\mathbf{x})$)
     - Only model decisions given the input examples (no model is constructed over the input examples)

Generative approach to classification

- We can model each class conditional population with a multivariate normal (Gaussian) distribution

$$\begin{aligned} \mathbf{x} &\sim N(\mu_1, \Sigma_1), \;\; y = 1 \\ \mathbf{x} &\sim N(\mu_0, \Sigma_0), \;\; y = 0 \end{aligned}$$

  where

$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \; \exp\{ -\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) \}$$

- How do we make decisions?

Mixture classifier cont'd

- Examples $\mathbf{x}$ are classified on the basis of which Gaussian explains the data better

$$\log \frac{P(\mathbf{x}|\mu_1,\Sigma_1)}{P(\mathbf{x}|\mu_0,\Sigma_0)} \begin{array}{ll} > & 0 \;\; y=1 \\ \leq & 0 \;\; y=0 \end{array}$$

or, more generally, when the classes have different a priori probabilities, we use the *posterior probability*

$$P(y=1|\mathbf{x}) = \frac{P(\mathbf{x}|\mu_1,\Sigma_1)P(y=1)}{P(\mathbf{x}|\mu_1,\Sigma_1)P(y=1) + P(X|\mu_0,\Sigma_0)P(y=0)}$$

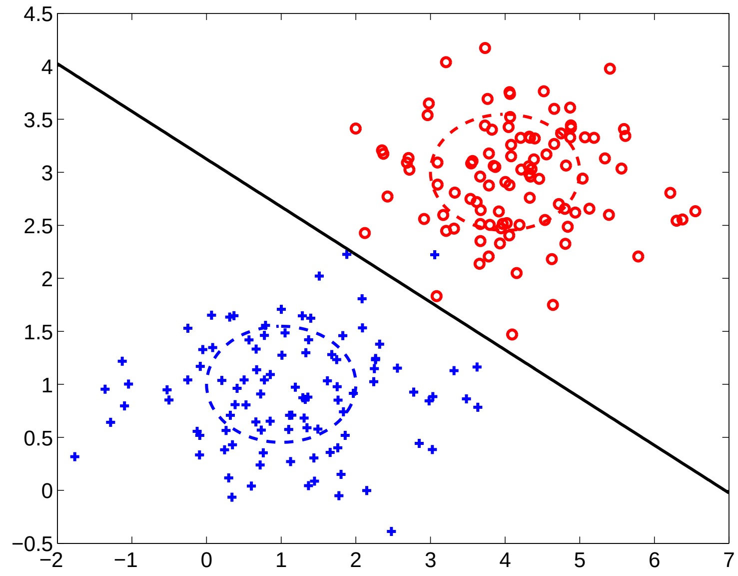
- The corresponding decision boundaries are

$$\log \frac{P(\mathbf{x}|\mu_1,\Sigma_1)}{P(\mathbf{x}|\mu_0,\Sigma_0)} = 0 \quad \text{or} \quad P(y=1|\mathbf{x}) = 0.5$$

Mixture classifier: decision rule

- Equal covariances

$$
\begin{aligned}
\mathbf{x} &\sim N(\mu_1, \Sigma), \quad y = 1 \\
\mathbf{x} &\sim N(\mu_0, \Sigma), \quad y = 0
\end{aligned}
$$

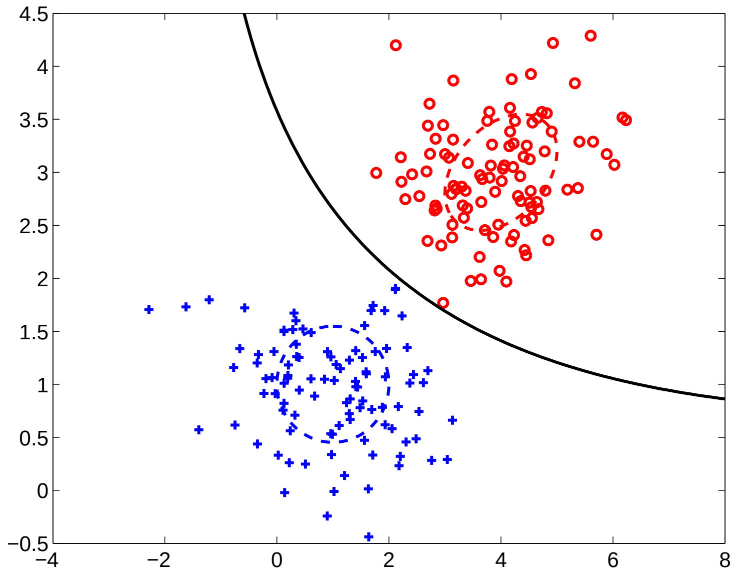


- The decision rule is *linear*

Mixture classifier: decision rule

- Unequal covariances

$$\begin{aligned} \mathbf{x} &\sim N(\mu_1, \Sigma_1), \;\; y = 1 \\ \mathbf{x} &\sim N(\mu_0, \Sigma_0), \;\; y = 0 \end{aligned}$$



- The decision rule is *quadratic*

Maximum likelihood estimation

- We can estimate the class conditional distributions $p(\mathbf{x}|\mu, \Sigma)$ separately (why?)

- For a multivariate Gaussian model

$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \ \text{exp}\{ -\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \}$$

given a random sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the maximum likelihood estimates of the parameters are:

1. Sample mean

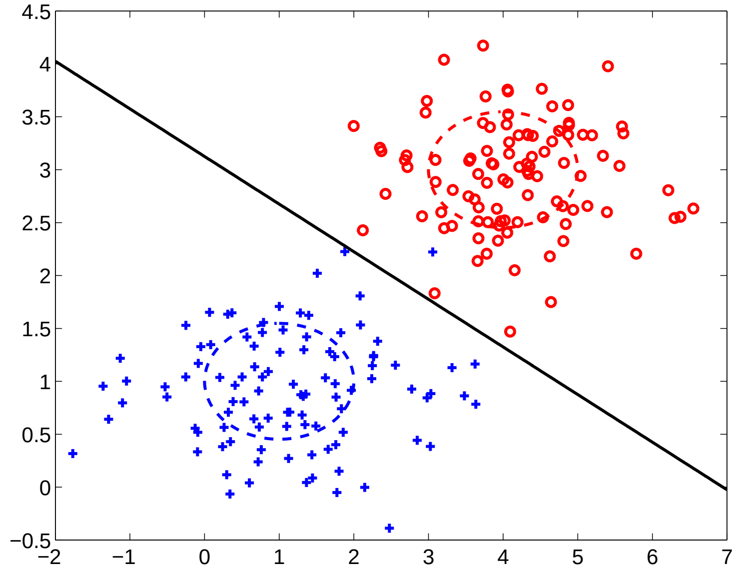
$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

2. Sample covariance

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \widehat{\mu})(\mathbf{x}_i - \widehat{\mu})^T$$

Discriminative classification

- If we are only interested in the classification decisions, why should we bother with a model over the input examples?



- We could directly estimate the *conditional distribution* of labels given the examples or $P(y|\mathbf{x}, \theta)$ where $\theta = \{\mu_0, \mu_1, \Sigma_0, \Sigma_1\}$.

- What do we gain? What do we loose?

Back to the Gaussians... (1-dim)

- When the classes are equally likely *a priori*, the posterior probabilily of the label $y = 1$ given x is given by

$$P(y = 1|x, \theta) = \frac{P(x|\mu_1, \sigma_1^2)}{P(x|\mu_1, \sigma_1^2) + P(x|\mu_0, \sigma_0^2)} = \frac{1}{1 + \exp\left\{ -\log \frac{P(x|\mu_1, \sigma_1^2)}{P(x|\mu_0, \sigma_0^2)} \right\}}$$

where $\theta = \{\mu_0, \mu_1, \sigma_1^2, \sigma_0^2\}$.

Back to the Gaussians... (1-dim)

- When the classes are equally likely *a priori*, the posterior probabilily of the label $y = 1$ given x is given by

$$P(y = 1|x, \theta) = \frac{P(x|\mu_1, \sigma_1^2)}{P(x|\mu_1, \sigma_1^2) + P(x|\mu_0, \sigma_0^2)} = \frac{1}{1 + \exp\left\{- \log \frac{P(x|\mu_1, \sigma_1^2)}{P(x|\mu_0, \sigma_0^2)}\right\}}$$

  where $\theta = \{\mu_0, \mu_1, \sigma_1^2, \sigma_0^2\}$.

- Since the decision boundary is *linear* or *quadratic*, we know that

$$\log \frac{P(x|\mu_1, \sigma_1^2)}{P(x|\mu_0, \sigma_0^2)} = \begin{cases} w_0 + w_1 x, \text{ when } \sigma_1^2 = \sigma_0^2 \\ w_0' + w_1' x + w_2' x^2, \text{ otherwise} \end{cases}$$

  for some coefficients w.

Back to the Gaussians... (1-dim)

- When the classes are equally likely *a priori*, the posterior probabilily of the label $y = 1$ given x is given by

$$P(y=1|x,\theta) = \frac{P(x|\mu_1,\sigma_1^2)}{P(x|\mu_1,\sigma_1^2)+P(x|\mu_0,\sigma_0^2)} = \frac{1}{1+\exp\left\{-\log\frac{P(x|\mu_1,\sigma_1^2)}{P(x|\mu_0,\sigma_0^2)}\right\}}$$

where $\theta = \{\mu_0, \mu_1, \sigma_1^2, \sigma_0^2\}$.

- Since the decision boundary is *linear* or *quadratic*, we know that

$$\log\frac{P(x|\mu_1,\sigma_1^2)}{P(x|\mu_0,\sigma_0^2)} = \begin{cases} w_0 + w_1 x, \text{ when } \sigma_1^2 = \sigma_0^2 \\ w_0' + w_1' x + w_2' x^2, \text{ otherwise} \end{cases}$$

for some coefficients w.

- When the variances are equal, we can write the posterior probability as a *squashed linear prediction*:

$$P(y=1|x,\mathbf{w}) = \frac{1}{1+\exp\{-(w_0+w_1 x)\}} = g\,(\,w_0 + w_1 x\,)$$

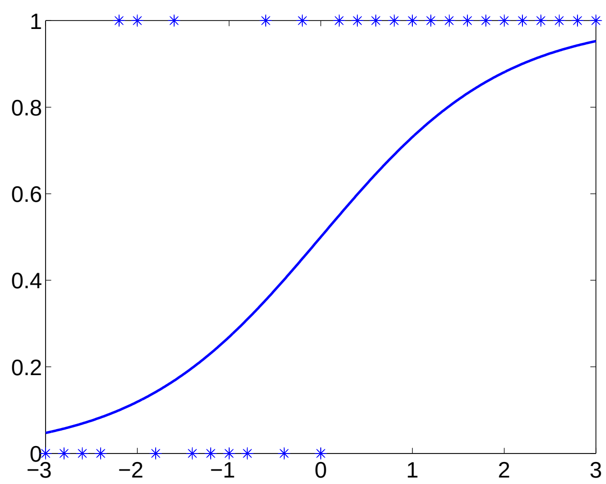
where $g(z) = (1+\exp\{-z\})^{-1}$.

Generalized linear models

- When the two Gaussian distributions have equal covariances, the posterior class probability $P(y = 1|\mathbf{x})$ from the mixture model reduces to a *logistic regression model*

$$P(y = 1|\mathbf{x}, \mathbf{w}) = g\,(\,w_0 + w_1 x_1 + \ldots + w_d x_d\,)$$

  where the parameters $\mathbf{w}$ are functions of μ_1,μ_0, and the common covariance Σ. Here $g(z) = (1 + \exp(-z))^{-1}$ is known as the *logistic function*.



- Robustness

Fitting logistic regression models

- Since the classification model gives a probability distribution over the labels y given the input $\mathbf{x}$ we can fit these models using the maximum likelihood criterion

  $$L(D; \mathbf{w}) = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w})$$

  where

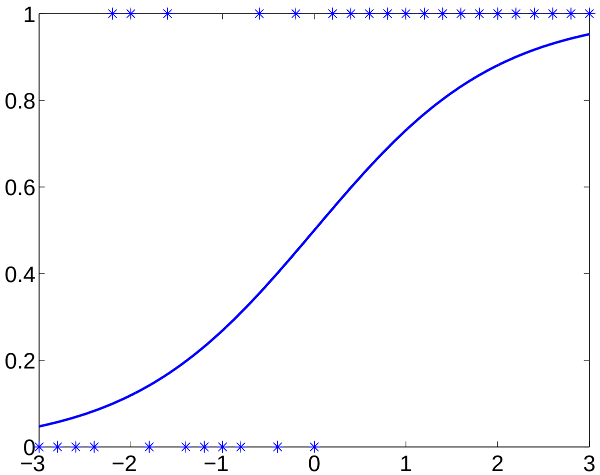
  $$P(y = 1|\mathbf{x}, \mathbf{w}) = g\,(\,w_0 + w_1 x_1 + \ldots + w_d x_d\,)$$

  Note: this is very different from the generative maximum likelihood fitting of mixture models

Stochastic gradient ascent for logistic regression

- We can try to maximize the likelihood in an *on-line* or incremental fashion.

  Given each training example $\mathbf{x}_i$ and the corresponding binary (0/1) label y_i, we change the parameters slightly to increase the (log-)probability of this particular label:



$$\begin{aligned}\mathbf{w} &\leftarrow \mathbf{w} + \epsilon \frac{\partial}{\partial \mathbf{w}} \log P(y_i|\mathbf{x}_i, \mathbf{w}) \\ &= \cdots \\ &= \mathbf{w} + \epsilon \underbrace{\Big(y_i - P(y_i = 1|\mathbf{x}_i, \mathbf{w})\Big)}_{\text{prediction error}} \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}\end{aligned}$$

  where ϵ is the *learning rate*.

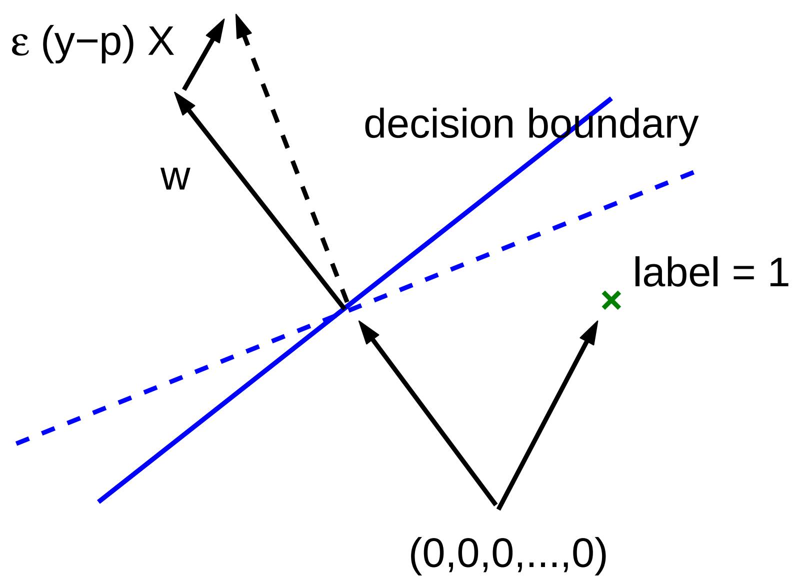
Stochastic gradient ascent cont'd

- Logistic regression model

$$P(y = 1|\mathbf{x}, \mathbf{w}) = g\left(w_0 + w_1 x_1 + \ldots + w_d x_d \right)$$

- Simple on-line parameter update rule

$$\mathbf{w} \;\leftarrow\; \mathbf{w} + \epsilon \underbrace{\Big(y_i - P(y_i = 1|\mathbf{x}_i, \mathbf{w})\Big)}_{\text{prediction error}} \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}$$



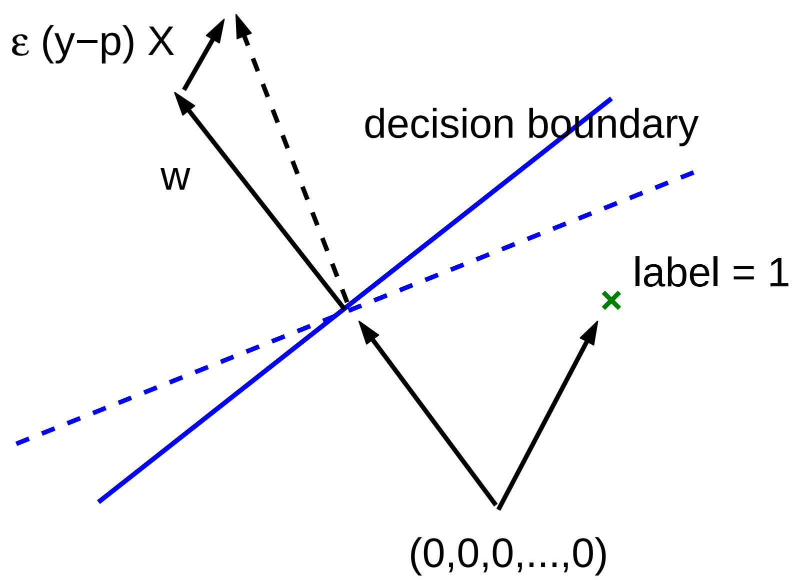
Stochastic gradient ascent: convergence

- The on-line learning method *converges* when we do not move in any direction on average:

$$\sum_{i=1}^{n} \underbrace{\left(y_i - P(y_i = 1|\mathbf{x}_i, \mathbf{w})\right)}_{\text{prediction error}} \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} = 0$$

where the summation is over the training set.

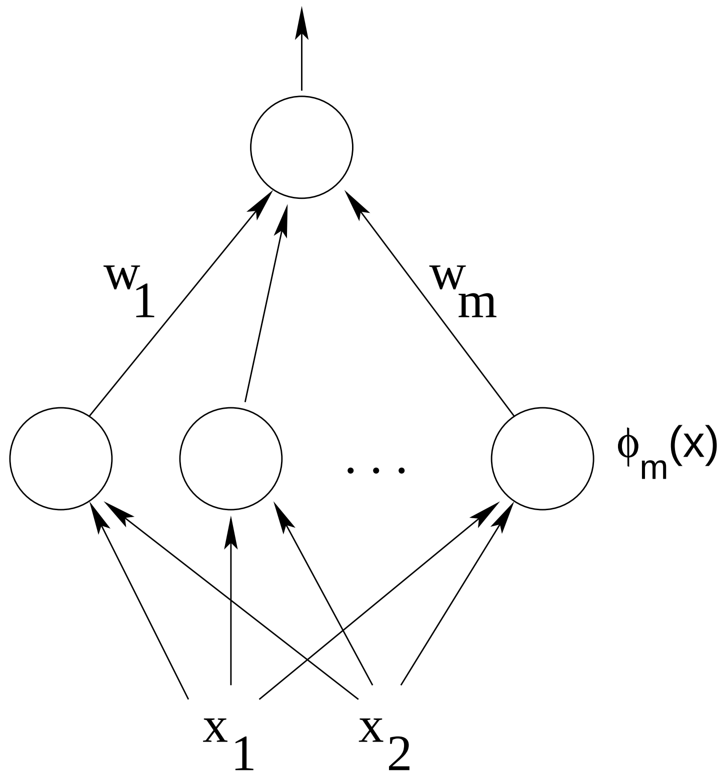
- The prediction error is again decorrelated with the inputs!



Additive models and classification

- Similarly to linear regression models, we can extend the logistic regression models via additive models

$$P(y = 1|\mathbf{x}, \mathbf{w}) = g\,(\,w_0 + w_1\phi_1(\mathbf{x}) + \ldots w_m\phi_m(\mathbf{x})\,)$$



- How should we then choose the basis functions $\phi_i(\mathbf{x})$?
- One approach is to make them adjustable...