Lecture 6: logistic regression

TTIC 31020: Introduction to Machine Learning

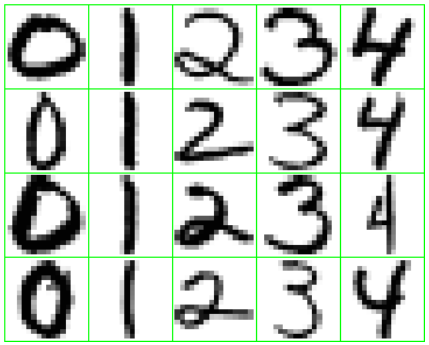
Instructors: Greg Shakhnarovich
Suriya Gunasekar

TTI-Chicago

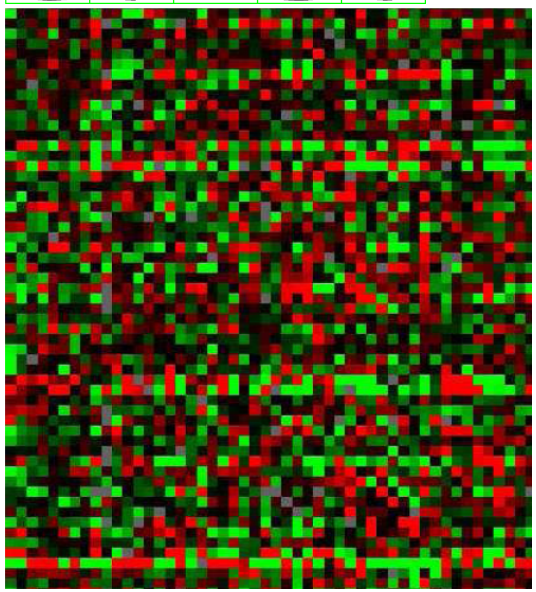
October 13, 2016

Classification

- Shifting gears: classification. Many successful applications of ML: vision, speech, medicine, etc.
- Setup: need to map $\mathbf{x} \in \mathcal{X}$ to a *label* $y \in \mathcal{Y}$.
- Examples:



digits recognition;
$\mathcal{Y} = \{0, \ldots, 9\}$



prediction from microarray data;
$\mathcal{Y} = \{\text{desease present/absent}\}$

Classification as regression

- Suppose we have a binary problem, $y \in \{-1, 1\}$
- Idea: treat it as regression, with squared loss
- Assuming the standard model $y = f(\mathbf{x}; \mathbf{w}) + \nu$, and solving with least squares, we get $\hat{\mathbf{w}}$.
- This corresponds to squared loss as a measure of classification performance! Does this make sense?

Classification as regression

- Suppose we have a binary problem, $y \in \{-1, 1\}$
- Idea: treat it as regression, with squared loss
- Assuming the standard model $y = f(\mathbf{x}; \mathbf{w}) + \nu$, and solving with least squares, we get $\hat{\mathbf{w}}$.
- This corresponds to squared loss as a measure of classification performance! Does this make sense?
- How do we decide on the label based on $f(\mathbf{x}; \hat{\mathbf{w}})$?

Classification as regression

$$f(\mathbf{x}; \hat{\mathbf{w}}) \;=\; w_0 + \hat{\mathbf{w}} \cdot \mathbf{x}$$

- Can't just take $\hat{y} = f(\mathbf{x}; \hat{\mathbf{w}})$ since it won't be a valid label.
- A reasonable *decision rule*:

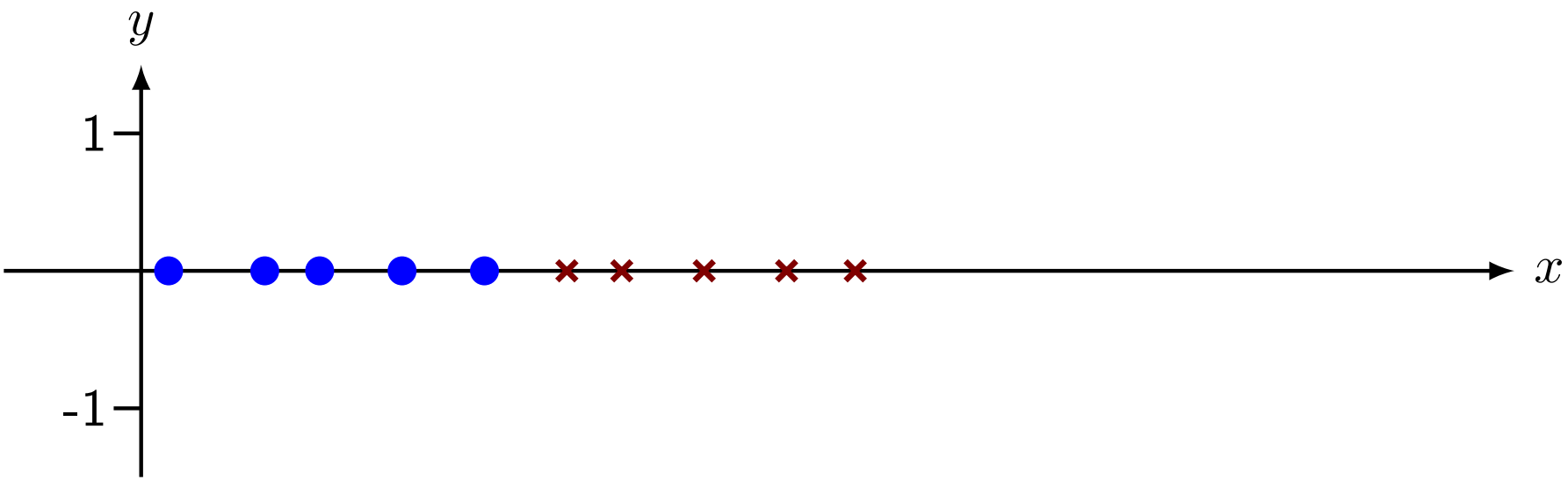
$$\text{decide on } \hat{y} = 1 \text{ if } f(\mathbf{x}; \hat{\mathbf{w}}) \geq 0, \text{ otherwise } \hat{y} = -1.$$

$$\hat{y} \;=\; \text{sign}\left(w_0 + \hat{\mathbf{w}} \cdot \mathbf{x}\right)$$

- This specifies a *linear classifier*:
  - The linear *decision boundary* (hyperplane) given by the equation $w_0 + \hat{\mathbf{w}} \cdot \mathbf{x} = 0$ separates the space into two "half-spaces".

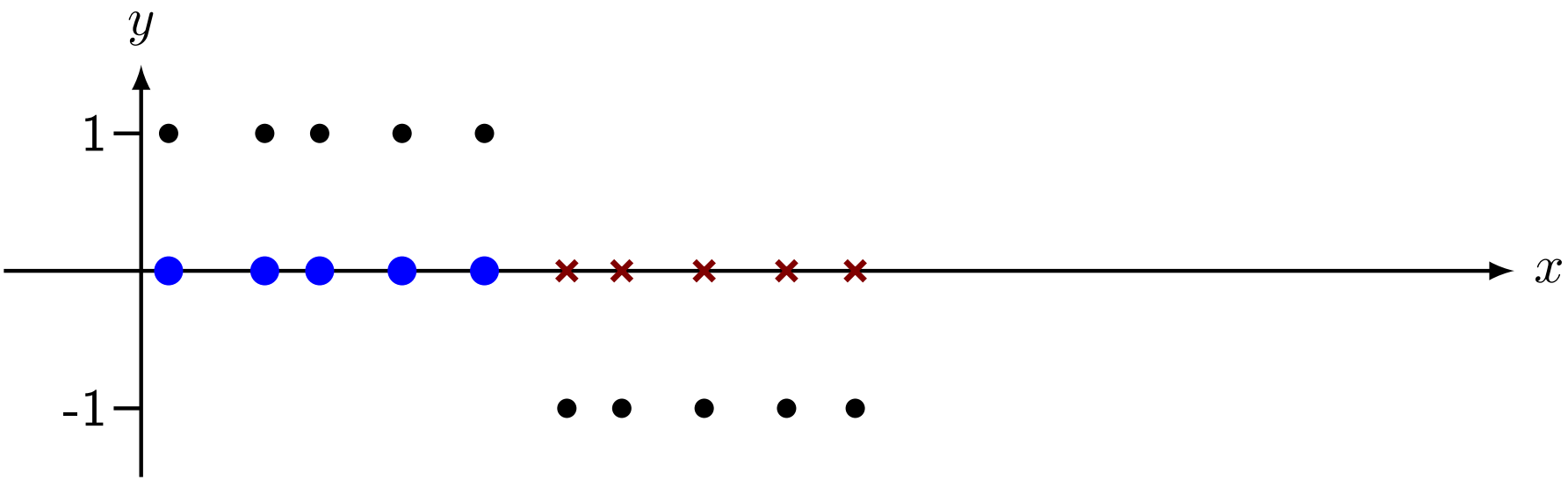
Classification as regression: example

- A 1D example:



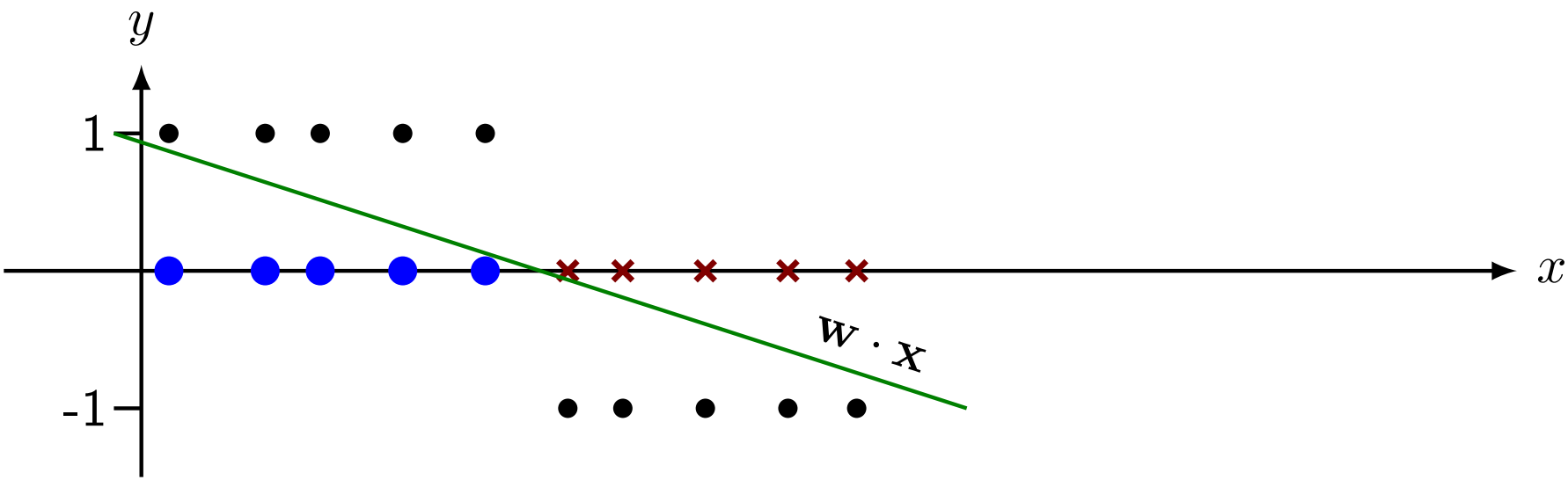
Classification as regression: example

- A 1D example:



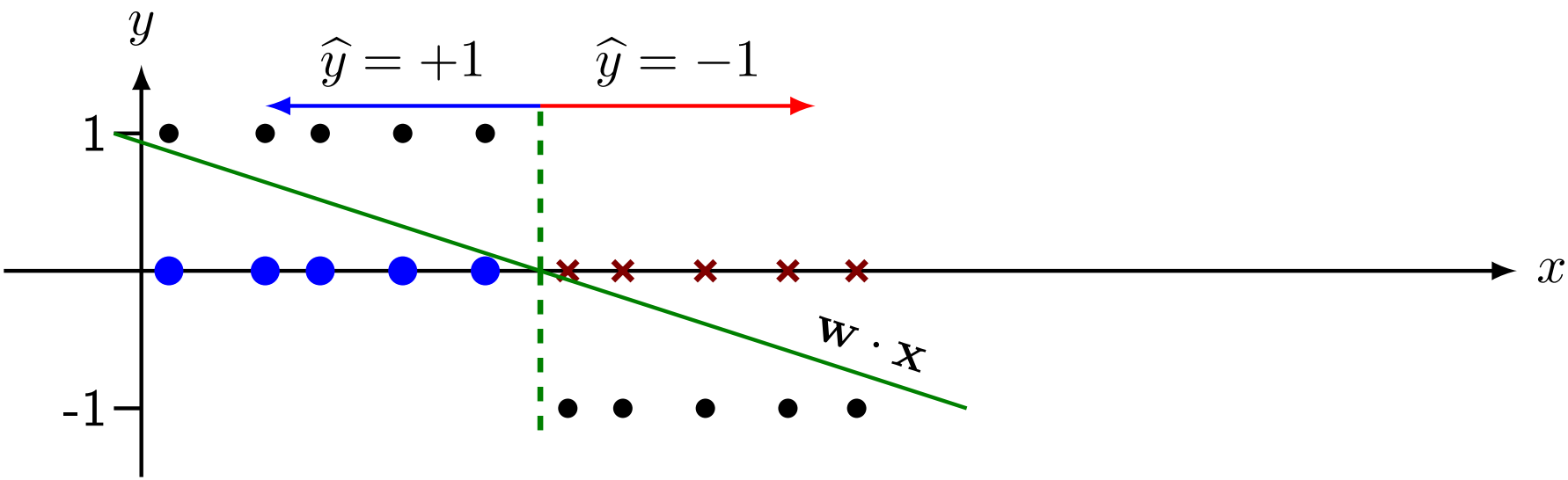
Classification as regression: example

- A 1D example:



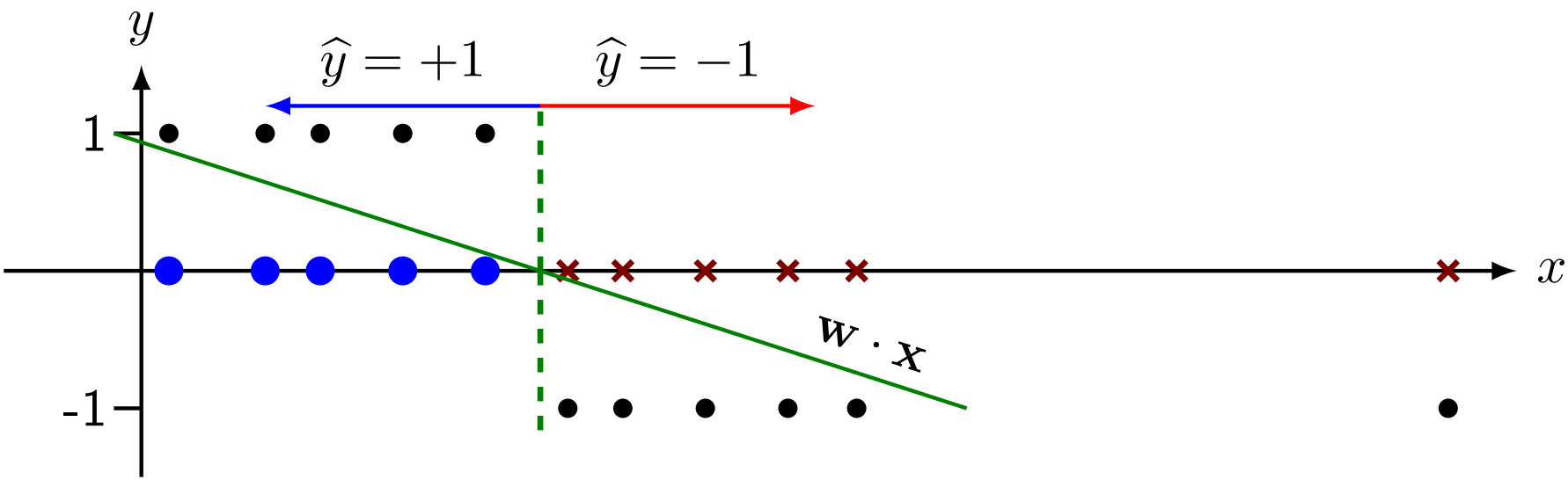
Classification as regression: example

- A 1D example:



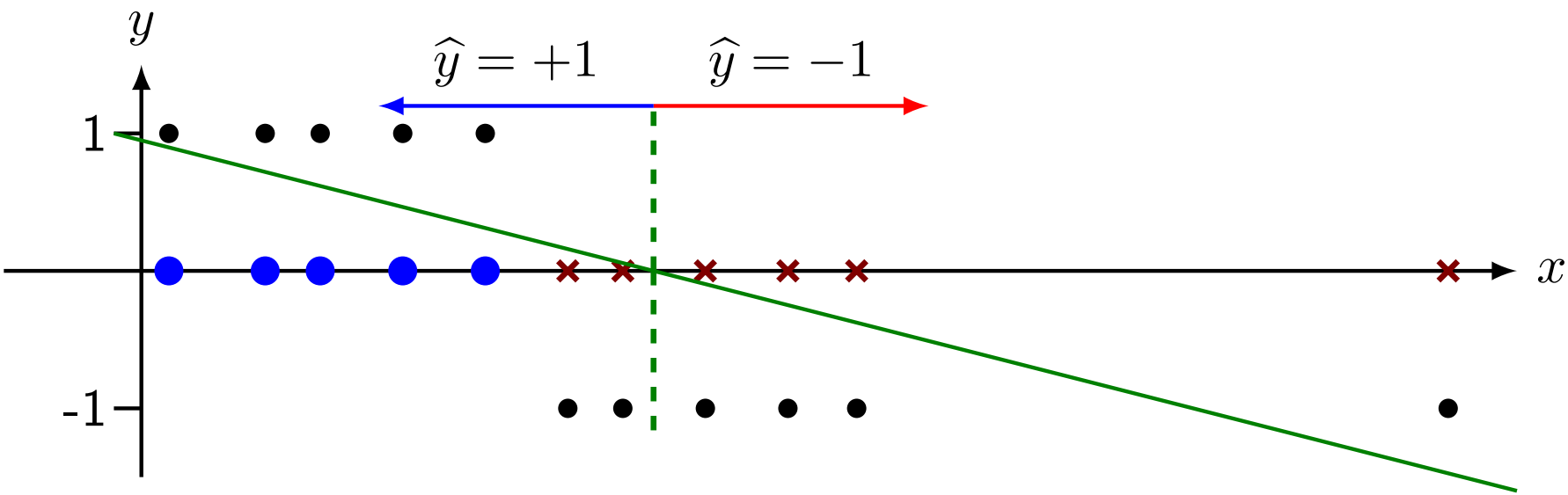
Classification as regression: example

- A 1D example:



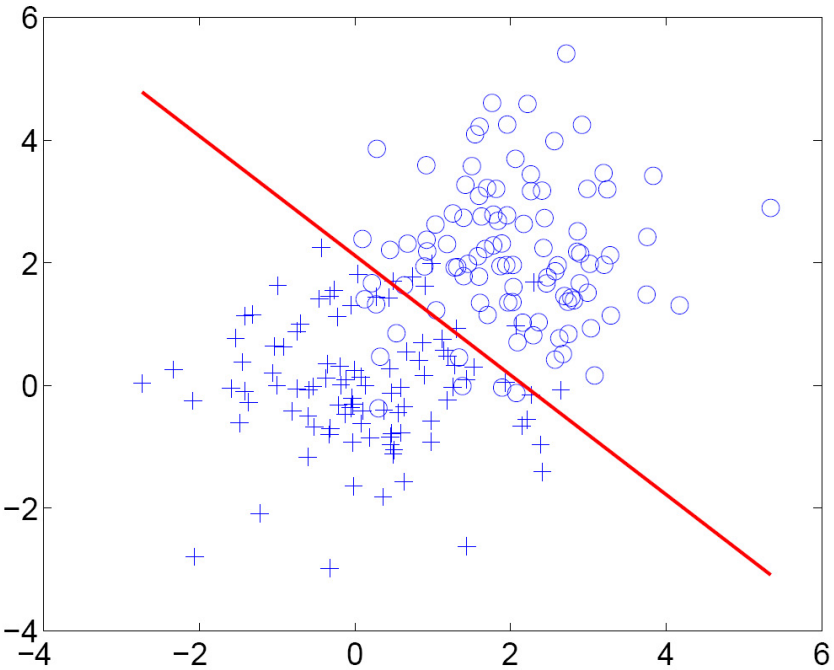
Classification as regression: example

- A 1D example:

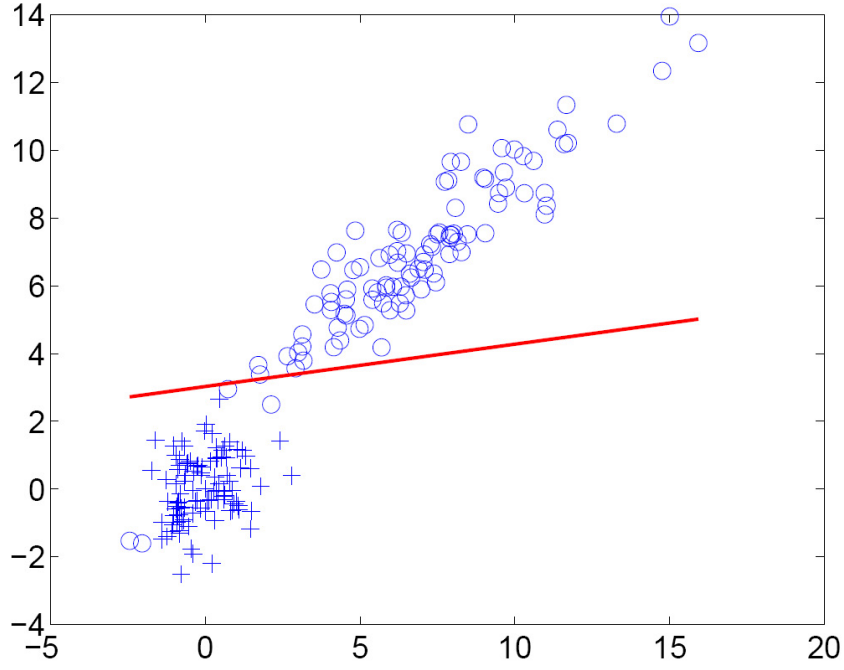


Classification as regression

- Same effect in 2D:

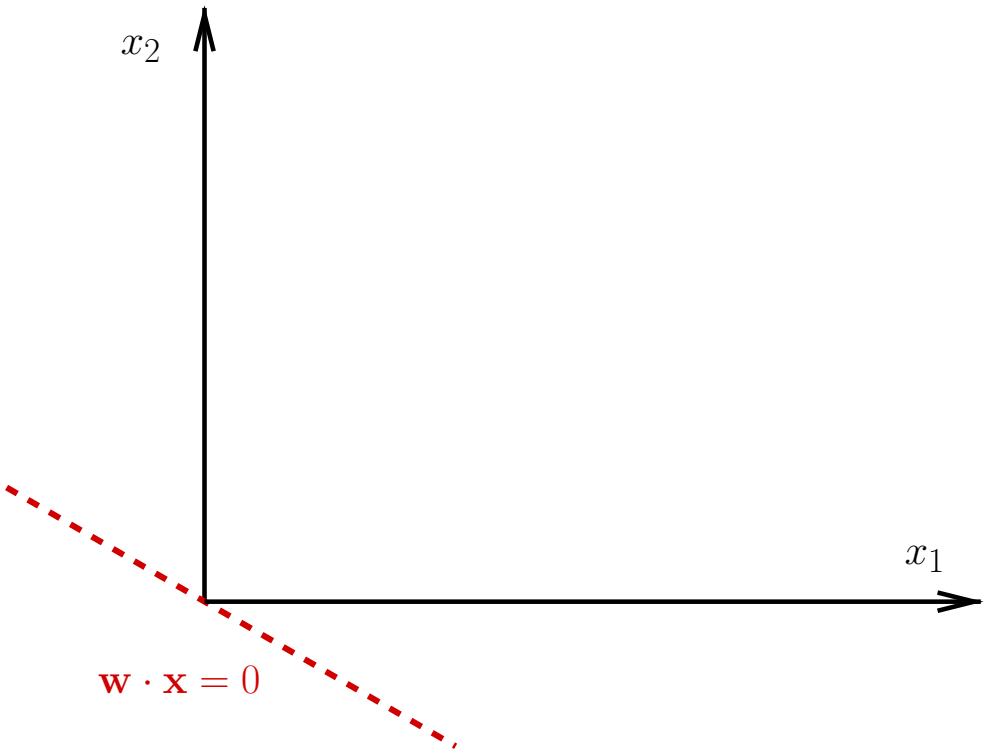


Seems to work well here

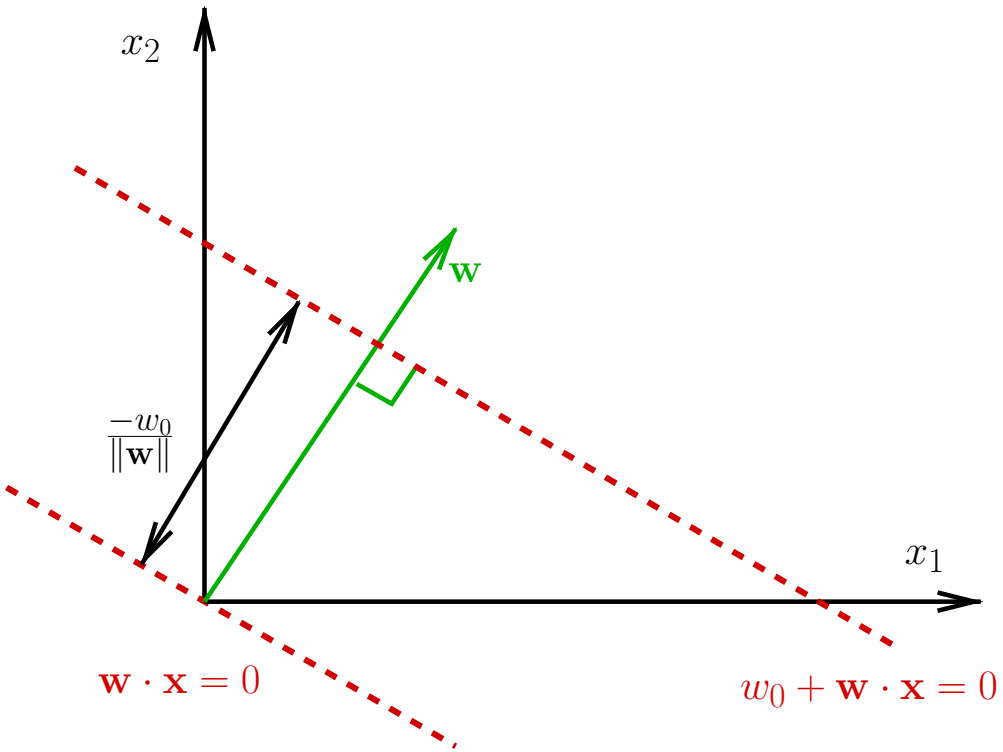


but not so well here

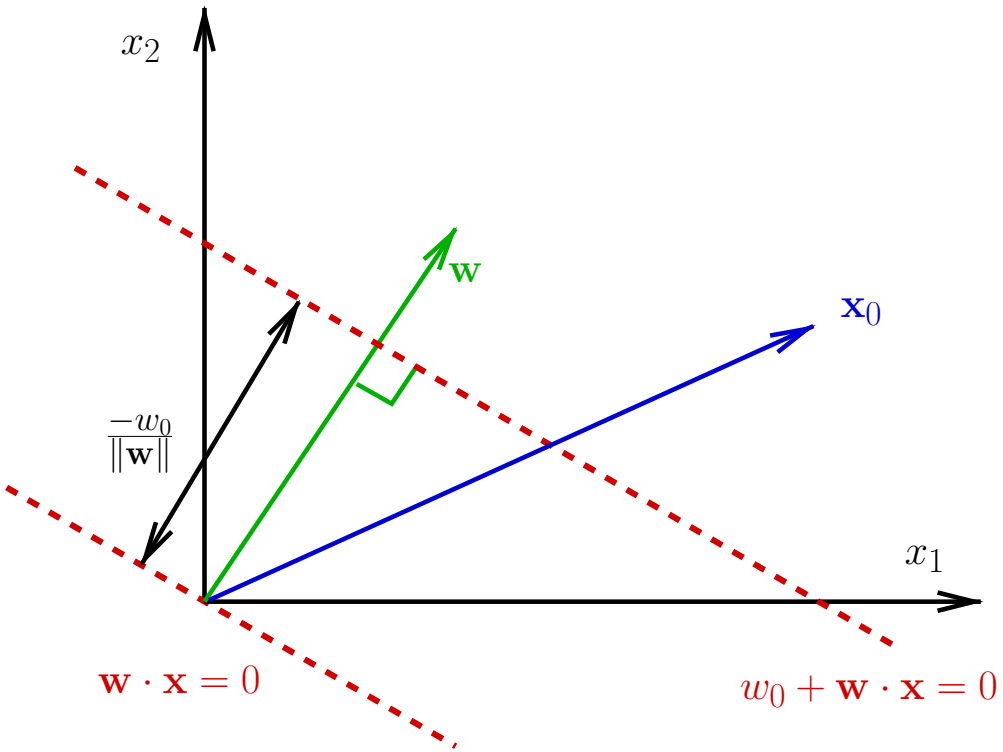
Geometry of projections



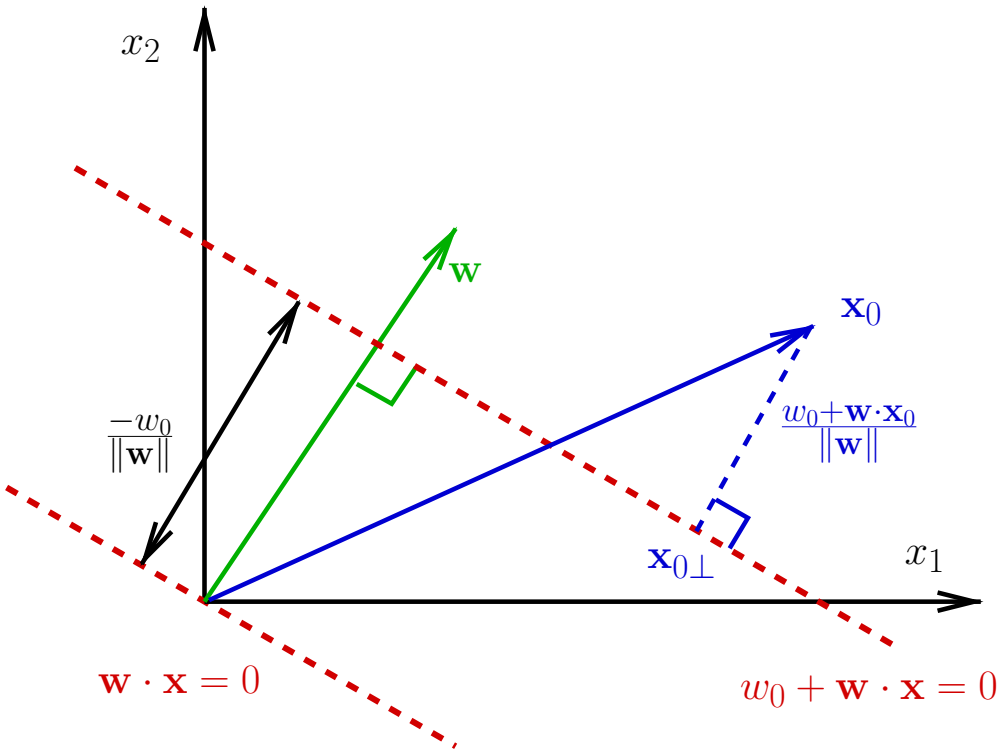
Geometry of projections



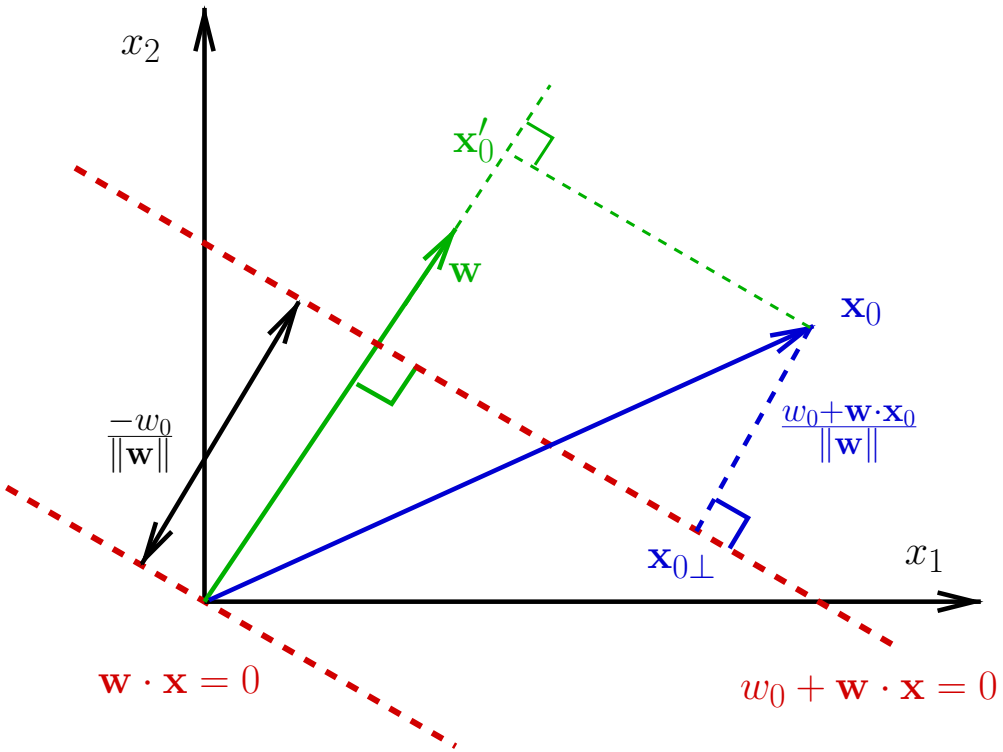
Geometry of projections



Geometry of projections

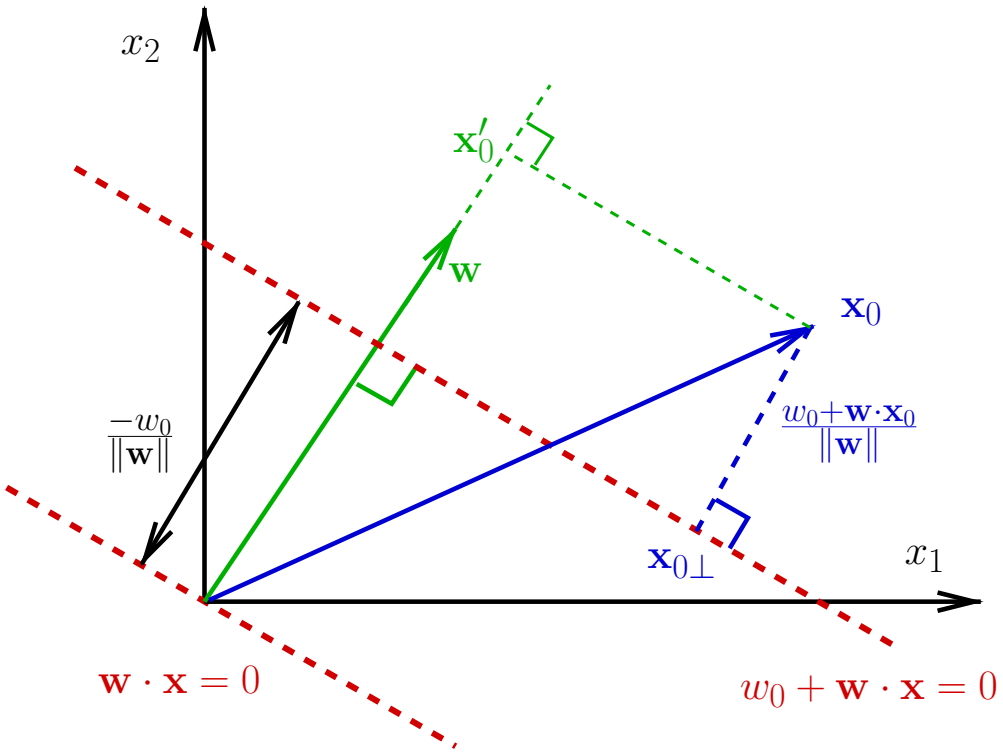


Geometry of projections



- $\mathbf{w}\cdot\mathbf{x}=0$: a line passing through the origin and *orthogonal* to $\mathbf{w}$
- $\mathbf{w}\cdot\mathbf{x}+w_0=0$ shifts the line along $\mathbf{w}$.

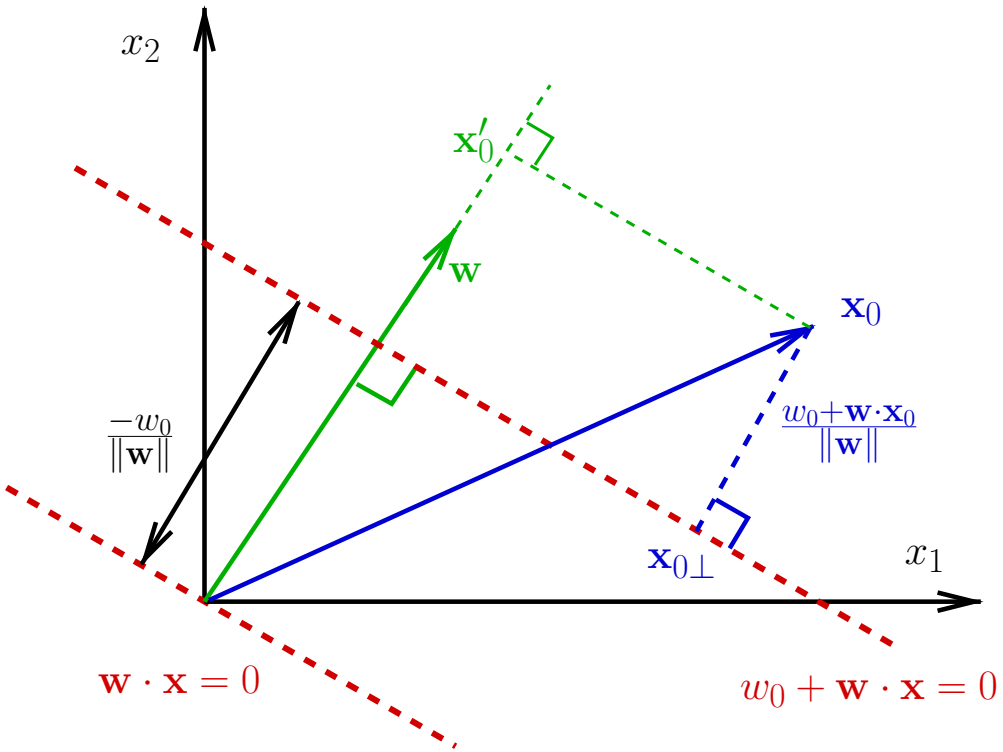
Geometry of projections



- $\mathbf{x}'$ is the projection of $\mathbf{x}$ on $\mathbf{w}$.

- $\mathbf{w}\cdot\mathbf{x}=0$: a line passing through the origin and *orthogonal* to $\mathbf{w}$
- $\mathbf{w}\cdot\mathbf{x}+w_0=0$ shifts the line along $\mathbf{w}$.

Geometry of projections



- $\mathbf{w}\cdot\mathbf{x}=0$: a line passing through the origin and *orthogonal* to $\mathbf{w}$
- $\mathbf{w}\cdot\mathbf{x}+w_0=0$ shifts the line along $\mathbf{w}$.

- $\mathbf{x}'$ is the projection of $\mathbf{x}$ on $\mathbf{w}$.
- Set up a new 1D coordinate system: $\mathbf{x} \rightarrow (w_0+\mathbf{w}\cdot\mathbf{x})/\|\mathbf{w}\|$.

Linear classifiers

$$\hat{y} = h(\mathbf{x}) \;=\; \text{sign}\,(w_0 + \mathbf{w} \cdot \mathbf{x})$$

- Classifying using a linear decision boundary effectively reduces the data dimension to 1.
- Need to find $\mathbf{w}$ (direction) and w_0 (location) of the boundary
- Want to minimize the expected zero/one loss for classifier $h : \mathcal{X} \to \mathcal{Y}$, which for $(\mathbf{x}, y)$ is

$$L(h(\mathbf{x}), y) \;=\; \begin{cases} 0 & \text{if } h(\mathbf{x}) = y, \\ 1 & \text{if } h(\mathbf{x}) \neq y. \end{cases}$$

Risk of a classifier

- The risk (expected loss) of a C-way classifier $h(\mathbf{x})$:

$$\begin{aligned} R(h) &= E_{\mathbf{x},y}\left[L(h(\mathbf{x}), y)\right] \\ &= \int_{\mathbf{x}} \sum_{c=1}^{C} L(h(\mathbf{x}), c)\, p(\mathbf{x}, y = c)\, d\mathbf{x} \end{aligned}$$

Risk of a classifier

- The risk (expected loss) of a C-way classifier $h(\mathbf{x})$:

$$\begin{aligned} R(h) &= E_{\mathbf{x},y}\left[L(h(\mathbf{x}), y)\right] \\ &= \int_{\mathbf{x}} \sum_{c=1}^{C} L(h(\mathbf{x}), c)\, p(\mathbf{x}, y = c)\, d\mathbf{x} \\ &= \int_{\mathbf{x}} \left[\sum_{c=1}^{C} L(h(\mathbf{x}), c)\, p\left(y = c \,|\, \mathbf{x}\right)\right] p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Risk of a classifier

- The risk (expected loss) of a C-way classifier $h(\mathbf{x})$:

$$\begin{aligned} R(h) &= E_{\mathbf{x},y}\left[L(h(\mathbf{x}), y)\right] \\ &= \int_{\mathbf{x}} \sum_{c=1}^{C} L(h(\mathbf{x}), c)\, p(\mathbf{x}, y = c)\, d\mathbf{x} \\ &= \int_{\mathbf{x}} \left[\sum_{c=1}^{C} L(h(\mathbf{x}), c)\, p\left(y = c \,|\, \mathbf{x}\right)\right] p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

- Clearly, it's enough to minimize the *conditional risk* for any $\mathbf{x}$:

$$R(h \,|\, \mathbf{x}) = \sum_{c=1}^{C} L(h(\mathbf{x}), c) p\left(y = c \,|\, \mathbf{x}\right).$$

Conditional risk of a classifier

$$R(h \,|\, \mathbf{x}) \;=\; \sum_{c=1}^{C} L(h(\mathbf{x}), c) p\left(y = c \,|\, \mathbf{x}\right)$$

Conditional risk of a classifier

$$
\begin{aligned}
R(h \,|\, \mathbf{x}) &= \sum_{c=1}^{C} L(h(\mathbf{x}), c) p\left(y = c \,|\, \mathbf{x}\right) \\
&= 0 \cdot p\left(y = h(\mathbf{x}) \,|\, \mathbf{x}\right) + 1 \cdot \sum_{c \neq h(\mathbf{x})} p\left(y = c \,|\, \mathbf{x}\right)
\end{aligned}
$$

Conditional risk of a classifier

$$
\begin{aligned}
R(h \,|\, \mathbf{x}) &= \sum_{c=1}^{C} L(h(\mathbf{x}), c) p\left(y = c \,|\, \mathbf{x}\right) \\
&= 0 \cdot p\left(y = h(\mathbf{x}) \,|\, \mathbf{x}\right) + 1 \cdot \sum_{c \neq h(\mathbf{x})} p\left(y = c \,|\, \mathbf{x}\right) \\
&= \sum_{c \neq h(\mathbf{x})} p\left(y = c \,|\, \mathbf{x}\right)
\end{aligned}
$$

Conditional risk of a classifier

$$
\begin{aligned}
R(h \,|\, \mathbf{x}) &= \sum_{c=1}^{C} L(h(\mathbf{x}), c) p\left(y = c \,|\, \mathbf{x}\right) \\
&= 0 \cdot p\left(y = h(\mathbf{x}) \,|\, \mathbf{x}\right) + 1 \cdot \sum_{c \neq h(\mathbf{x})} p\left(y = c \,|\, \mathbf{x}\right) \\
&= \sum_{c \neq h(\mathbf{x})} p\left(y = c \,|\, \mathbf{x}\right) = 1 - p\left(y = h(\mathbf{x}) \,|\, \mathbf{x}\right).
\end{aligned}
$$

- To minimize conditional risk given $\mathbf{x}$, the classifier must decide

$$
h(\mathbf{x}) = \operatorname*{argmax}_{c} p\left(y = c \,|\, \mathbf{x}\right).
$$

- This is the *best possible* classifier in terms of generalization, i.e. expected misclassification rate on new examples.

Log-odds ratio

- Optimal rule $h(\mathbf{x}) = \text{argmax}_c \, p\left(y = c \,|\, \mathbf{x}\right)$ is equivalent to

$$h(\mathbf{x}) = c^* \quad \Leftrightarrow \quad \frac{p\left(y = c^* \,|\, \mathbf{x}\right)}{p\left(y = c \,|\, \mathbf{x}\right)} \geq 1 \quad \forall c$$

Log-odds ratio

- Optimal rule $h(\mathbf{x}) = \text{argmax}_c \, p\left(y = c \,|\, \mathbf{x}\right)$ is equivalent to

$$h(\mathbf{x}) = c^* \quad \Leftrightarrow \quad \frac{p\left(y = c^* \,|\, \mathbf{x}\right)}{p\left(y = c \,|\, \mathbf{x}\right)} \geq 1 \quad \forall c$$

$$\Leftrightarrow \quad \log \frac{p\left(y = c^* \,|\, \mathbf{x}\right)}{p\left(y = c \,|\, \mathbf{x}\right)} \geq 0 \quad \forall c$$

- For the binary case,

$$h(\mathbf{x}) = 1 \quad \Leftrightarrow \quad \log \frac{p\left(y = 1 \,|\, \mathbf{x}\right)}{p\left(y = 0 \,|\, \mathbf{x}\right)} \geq 0.$$

The logistic model

- We can model the (unknown) decision boundary directly:

$$\log \frac{p\left(y=1 \,|\, \mathbf{x}\right)}{p\left(y=0 \,|\, \mathbf{x}\right)} \;=\; w_0 + \mathbf{w} \cdot \mathbf{x} \;=\; 0.$$

- Since $p\left(y=1 \,|\, \mathbf{x}\right) = 1 - p\left(y=0 \,|\, \mathbf{x}\right)$, we have (after exponentiating):

$$\frac{p\left(y=1 \,|\, \mathbf{x}\right)}{1 - p\left(y=1 \,|\, \mathbf{x}\right)} \;=\; \exp(w_0 + \mathbf{w} \cdot \mathbf{x}) \;=\; 1$$

The logistic model

- We can model the (unknown) decision boundary directly:

$$\log \frac{p\left(y=1 \,|\, \mathbf{x}\right)}{p\left(y=0 \,|\, \mathbf{x}\right)} \;=\; w_0 + \mathbf{w} \cdot \mathbf{x} \;=\; 0.$$

- Since $p\left(y=1 \,|\, \mathbf{x}\right) = 1 - p\left(y=0 \,|\, \mathbf{x}\right)$, we have (after exponentiating):

$$\frac{p\left(y=1 \,|\, \mathbf{x}\right)}{1 - p\left(y=1 \,|\, \mathbf{x}\right)} \;=\; \exp(w_0 + \mathbf{w} \cdot \mathbf{x}) \;=\; 1$$

$$\Rightarrow \frac{1}{p\left(y=1 \,|\, \mathbf{x}\right)} \;=\; 1 + \exp(-w_0 - \mathbf{w} \cdot \mathbf{x}) \;=\; 2$$

The logistic model

- We can model the (unknown) decision boundary directly:

$$\log \frac{p\left(y=1 \,|\, \mathbf{x}\right)}{p\left(y=0 \,|\, \mathbf{x}\right)} \;=\; w_0 + \mathbf{w} \cdot \mathbf{x} \;=\; 0.$$

- Since $p\left(y=1 \,|\, \mathbf{x}\right) = 1 - p\left(y=0 \,|\, \mathbf{x}\right)$, we have (after exponentiating):

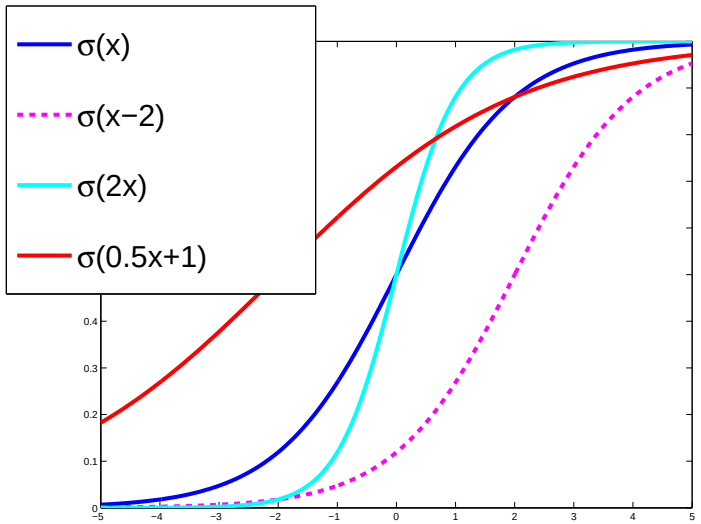
$$\begin{aligned}
\frac{p\left(y=1 \,|\, \mathbf{x}\right)}{1 - p\left(y=1 \,|\, \mathbf{x}\right)} \;&=\; \exp(w_0 + \mathbf{w} \cdot \mathbf{x}) \;=\; 1 \\
\Rightarrow \frac{1}{p\left(y=1 \,|\, \mathbf{x}\right)} \;&=\; 1 + \exp(-w_0 - \mathbf{w} \cdot \mathbf{x}) \;=\; 2 \\
\Rightarrow p\left(y=1 \,|\, \mathbf{x}\right) \;&=\; \frac{1}{1 + \exp(-w_0 - \mathbf{w} \cdot \mathbf{x})} \;=\; \frac{1}{2}.
\end{aligned}$$

The logistic function

$$p\left(y=1 \,|\, \mathbf{x}\right) \;=\; \frac{1}{1+\exp(-w_0 - \mathbf{w}\cdot\mathbf{x})}$$

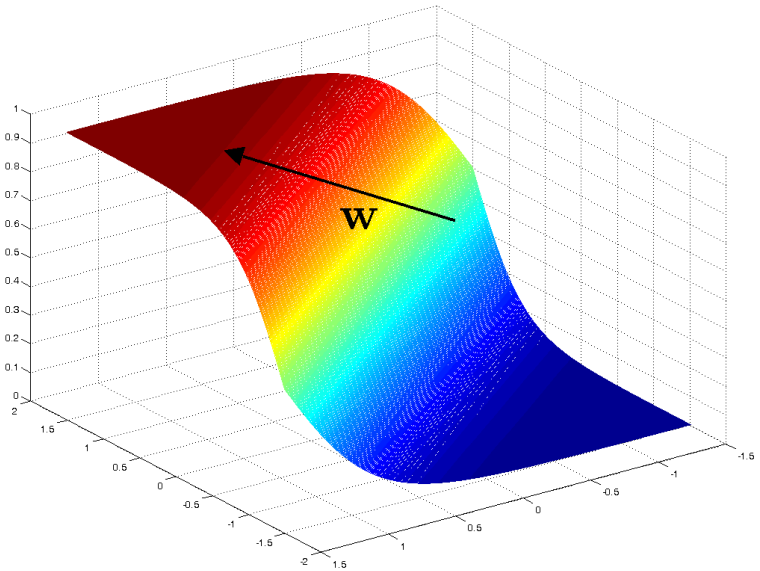
- The logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$:
  For any x, $0 \leq \sigma(x) \leq 1$;
  Monotonic, $\sigma(-\infty) = 0$, $\sigma(+\infty) = 1$

- $\sigma(0) = 1/2$. To shift the crossing to an arbitrary z: $\sigma(x - z)$.
- To change the "slope": $\sigma(ax)$.



Logistic function in $\mathbb{R}^d$

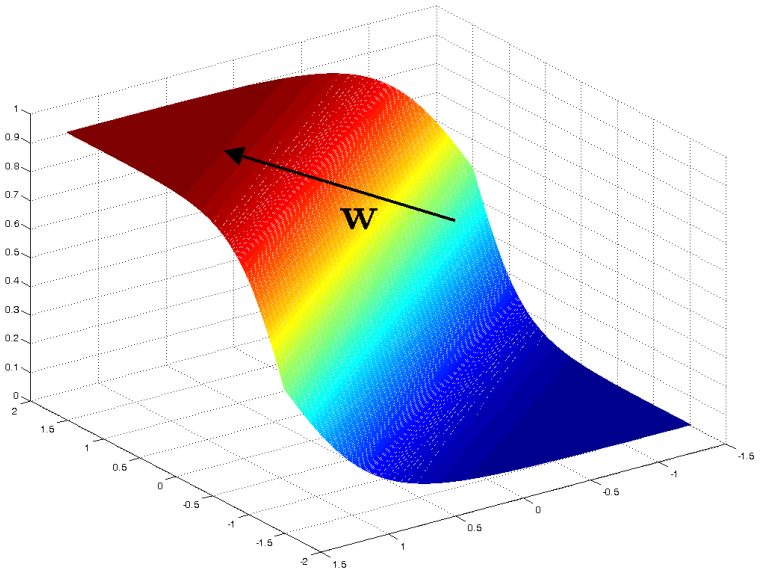
- What if $\mathbf{x} \in \mathbb{R}^d = [x_1 \dots x_d]$?
- $\sigma(w_0 + \mathbf{w} \cdot \mathbf{x})$ is a scalar function of a scalar variable $w_0 + \mathbf{w} \cdot \mathbf{x}$.



- the direction of $\mathbf{w}$ determines orientation;
- w_0 determines the location;

Logistic function in $\mathbb{R}^d$

- What if $\mathbf{x} \in \mathbb{R}^d = [x_1 \dots x_d]$?
- $\sigma(w_0 + \mathbf{w} \cdot \mathbf{x})$ is a scalar function of a scalar variable $w_0 + \mathbf{w} \cdot \mathbf{x}$.

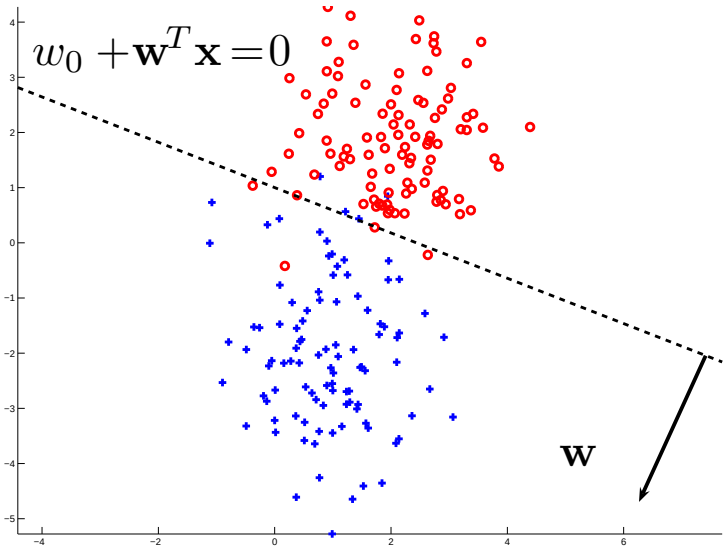


- the direction of $\mathbf{w}$ determines orientation;
- w_0 determines the location;
- $\|\mathbf{w}\|$ determines the slope.

Logistic regression: decision boundary

$$p\left(y=1\,|\,\mathbf{x}\right) \;=\; \sigma(w_0+\mathbf{w}\cdot\mathbf{x}) = 1/2 \;\Leftrightarrow\; w_0+\mathbf{w}\cdot\mathbf{x}=0$$

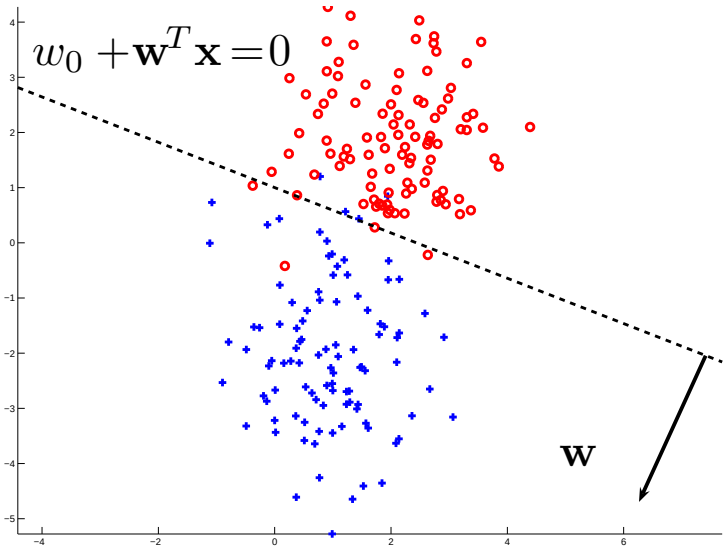
- With linear logistic model we get a linear decision boundary.

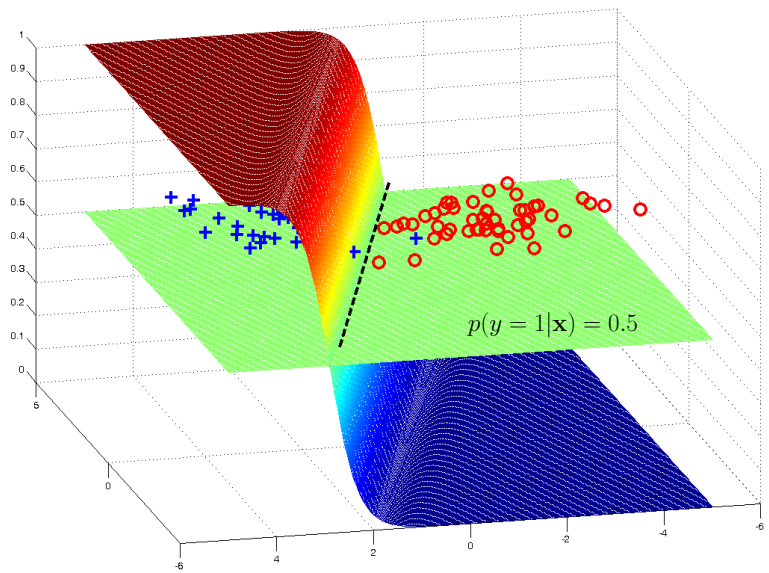


Logistic regression: decision boundary

$$p\left(y=1\,|\,\mathbf{x}\right) \;=\; \sigma(w_0+\mathbf{w}\cdot\mathbf{x}) = 1/2 \;\Leftrightarrow\; w_0+\mathbf{w}\cdot\mathbf{x}=0$$

- With linear logistic model we get a linear decision boundary.





Likelihood under the logistic model

- Regression: observe values, measure residuals under the model.
- Logistic regression: observe labels, measure their probability under the model.

$$p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}\right) = \begin{cases} \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i) & \text{if } y_i = 1, \\ 1 - \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i) & \text{if } y_i = 0 \end{cases}$$

Likelihood under the logistic model

- Regression: observe values, measure residuals under the model.
- Logistic regression: observe labels, measure their probability under the model.

$$
\begin{aligned}
p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}\right) &= \begin{cases} \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i) & \text{if } y_i = 1, \\ 1 - \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i) & \text{if } y_i = 0 \end{cases} \\
&= \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i)^{y_i} \left(1 - \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i)\right)^{1 - y_i}.
\end{aligned}
$$

Likelihood under the logistic model

- Regression: observe values, measure residuals under the model.
- Logistic regression: observe labels, measure their probability under the model.

$$p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}\right) = \begin{cases} \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i) & \text{if } y_i = 1, \\ 1 - \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i) & \text{if } y_i = 0 \end{cases}$$
$$= \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i)^{y_i} \left(1 - \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i)\right)^{1-y_i}.$$

- The log-likelihood of $\mathbf{w}$:

$$\log p(Y|X; \mathbf{w}) = \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i;\, \mathbf{w}\right)$$
$$= \sum_{i=1}^{N} y_i \log \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i) + (1 - y_i) \log\left(1 - \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i)\right)$$

The maximum likelihood solution

$$\log p(Y|X;\mathbf{w}) = \sum_{i=1}^{N} y_i \log \sigma(w_0 + \mathbf{w}\cdot\mathbf{x}_i) + (1-y_i)\log\left(1-\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i)\right)$$

- Setting the derivatives to zero, we get

$$\frac{\partial}{\partial w_0}\log p(Y|X;\mathbf{w}) = \sum_{i=1}^{N}\left(y_i - \sigma\left(w_0+\mathbf{w}\cdot\mathbf{x}_i\right)\right) = 0;$$

$$\frac{\partial}{\partial w_j}\log p(Y|X;\mathbf{w}) = \sum_{i=1}^{N}\left(y_i - \sigma\left(w_0+\mathbf{w}\cdot\mathbf{x}_i\right)\right)x_{ij} = 0.$$

- We can treat $y_i - p\left(y_i \,|\, \mathbf{x}_i\right) = y_i - \sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i)$ as the *prediction error* of the model on $\mathbf{x}_i, y_i$.
- As with linear regression: prediction errors are uncorrelated with any linear function of the data.

Gradient ascent

- We can cycle through the examples, accumulating the gradient, and then applying the accumulated value to form an update

$$\begin{aligned}\mathbf{w}_{new} &:= \mathbf{w} + \eta \frac{\partial}{\partial \mathbf{w}} \log p(X; \mathbf{w}) \\ &= \mathbf{w} + \eta \sum_{i=1}^{N} (y_i - \sigma(w_0 + \mathbf{w} \cdot \mathbf{x}_i)) \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}\end{aligned}$$

- Remember: need to choose η rather carefully:
  - Too small $\Rightarrow$ slow convergence;
  - Too large: $\Rightarrow$ overshoot and osillation.

Newton-Raphson

- The *Newton-Raphson* algorithm: approximate the local shape of $\log p$ as a quadratic function.

$$\mathbf{w}_{new} \;:=\; \mathbf{w} \;+\; \mathbf{H}^{-1} \frac{\partial}{\partial \mathbf{w}} \log p(X; \mathbf{w}),$$

where $\mathbf{H}$ is the *Hessian* matrix of second derivatives:

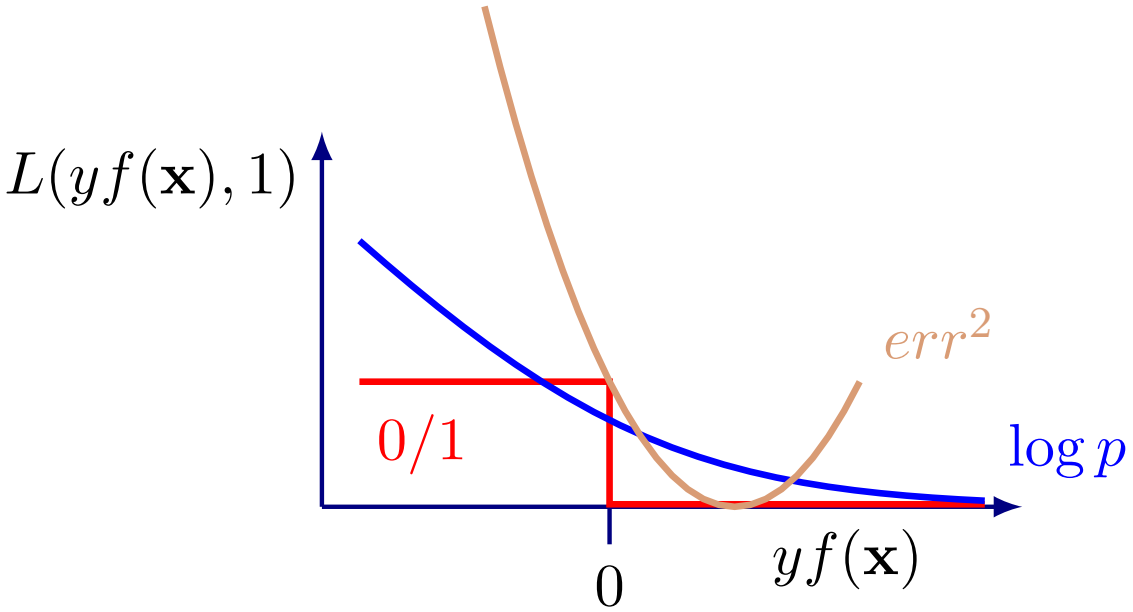
$$\mathbf{H} \;=\; \begin{bmatrix} \frac{\partial^2 \log p}{\partial w_0^2} & \frac{\partial^2 \log p}{\partial w_0 w_1} & \cdots & \frac{\partial^2 \log p}{\partial w_0 w_d} \\ \frac{\partial^2 \log p}{\partial w_0 w_1} & \frac{\partial^2 \log p}{\partial w_1^2} & \cdots & \frac{\partial^2 \log p}{\partial w_1 w_d} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 \log p}{\partial w_d w_0} & \frac{\partial^2 \log p}{\partial w_d w_1} & \cdots & \frac{\partial^2 \log p}{\partial w_d^2} \end{bmatrix}$$

Surrogate loss

- Recall that we really want to minimize 0/1 loss
- Instead, we are minimizing the log-loss:

$$\underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i;\, \mathbf{w}) \;=\; \underset{\mathbf{w}}{\operatorname{argmin}} - \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i\,;\mathbf{w})$$

- This is a *surrogate* loss; we work with it since it is not computationally feasible to optimize the 0/1 loss directly.



Generalized additive models

- As with regression we can extend this framework to arbitrary features (basis functions):

$$p\left(y=1\,|\,\mathbf{x}\right) \;=\; \sigma\left(w_0 + \phi_1(\mathbf{x}) + \ldots + \phi_m(\mathbf{x})\right).$$

- Example: quadratic logistic regression in 2D

$$p\left(y=1\,|\,\mathbf{x}\right) \;=\; \sigma\left(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2\right).$$

Generalized additive models

- As with regression we can extend this framework to arbitrary features (basis functions):

$$p\left(y=1\,|\,\mathbf{x}\right) \;=\; \sigma\left(w_0+\phi_1(\mathbf{x})+\ldots+\phi_m(\mathbf{x})\right).$$

- Example: quadratic logistic regression in 2D

$$p\left(y=1\,|\,\mathbf{x}\right) \;=\; \sigma\left(w_0+w_1x_1+w_2x_2+w_3x_1^2+w_4x_2^2\right).$$

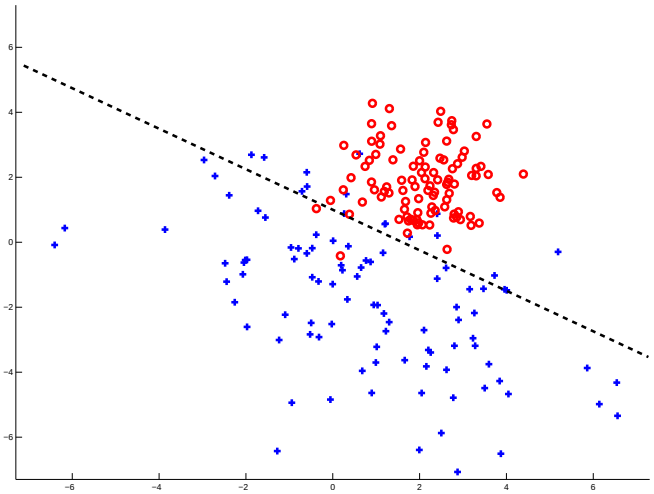
  - Decision boundary of this classifier:

$$w_0+w_1x_1+w_2x_2+w_3x_1^2+w_4x_2^2=0,$$

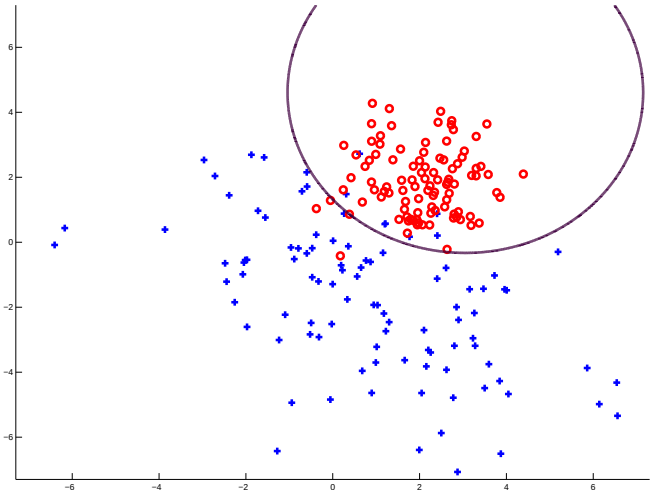
  i.e. it's a quadratic decision boundary.

Logistic regression: 2D example

Linear

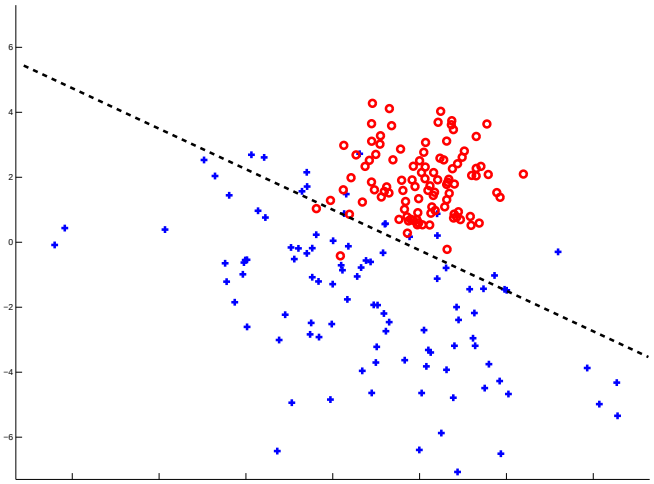


Quadratic

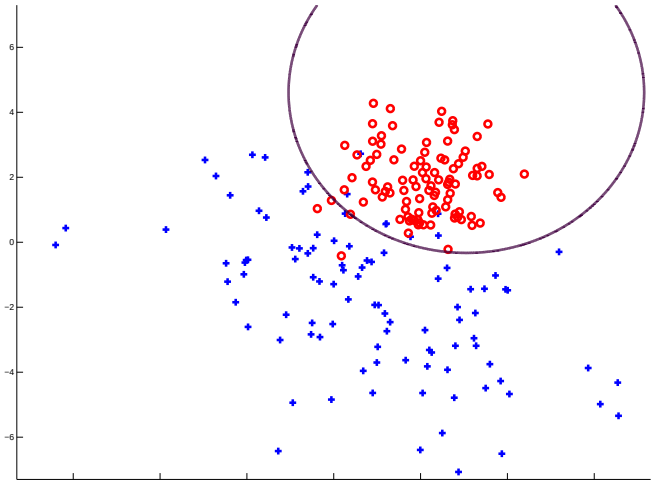


Logistic regression: 2D example

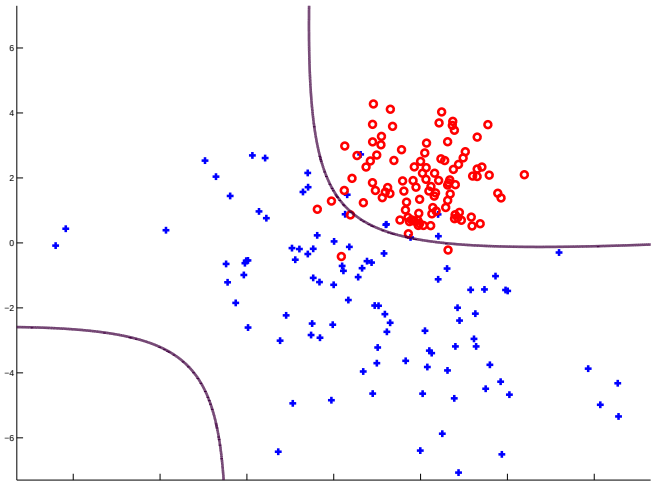
Linear



Quadratic



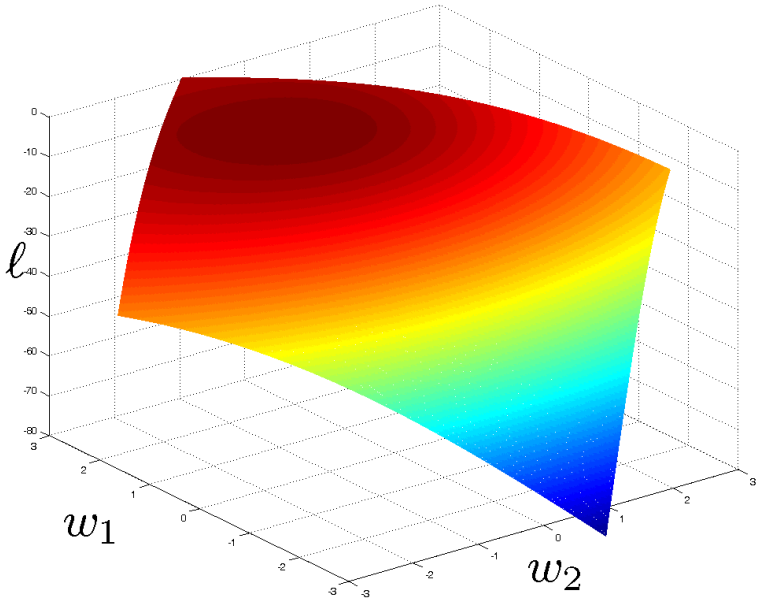
We can also include $x_1 x_2$:



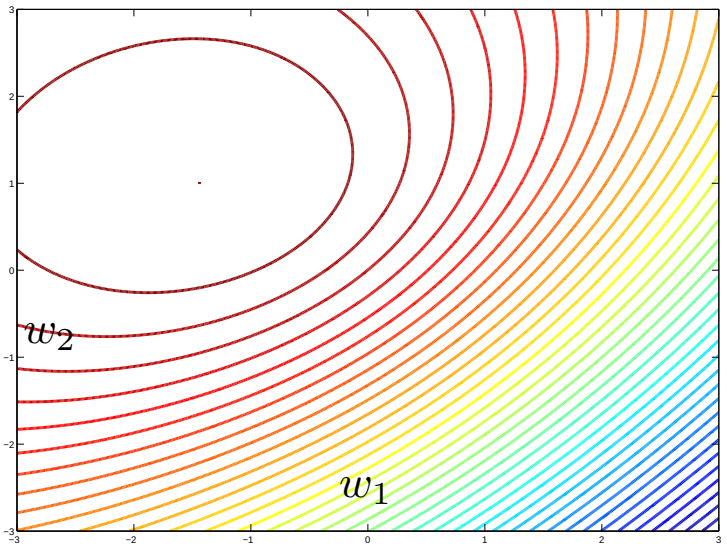
Visualizing the log-likelihood surface

- We will look at a 2D example, and assume $w_0 = 0$, i.e. our model will be $\hat{p}(y = 1|\mathbf{x}) = \sigma(w_1 x_1 + w_2 x_2)$.

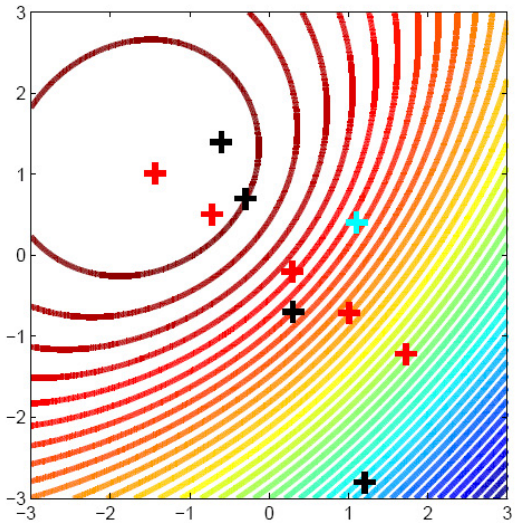
$\log p$ as a function of $\mathbf{w}$

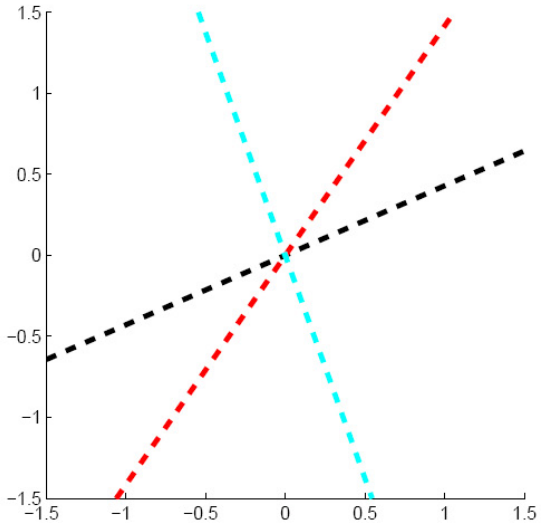


Contour plot: high/low

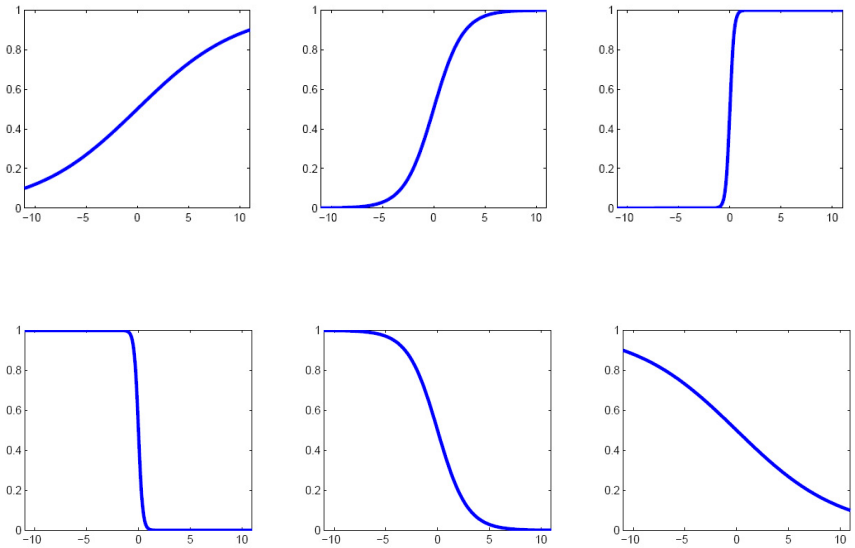


Mapping from boundaries to $\mathbf{w}$





- A line $\alpha\mathbf{w}$ in the parameter space $\Leftrightarrow$ identical decision boundaries of the form $\alpha\mathbf{w} \cdot \mathbf{x} = 0$.
- The sign of α determines the direction.
- Think about the effect of w_0



Overfitting with logistic regression

- We can get the same decision boundary with an infinite number of settings for $\mathbf{w}$.
- When the data are *separable* by $w_0 + \alpha \mathbf{w} \cdot \mathbf{x} = 0$, what's the best choice for α?

$$p(y = 1 \,|\, \mathbf{x}) \;=\; \sigma(w_0 + \alpha \mathbf{w} \cdot \mathbf{x}).$$

Overfitting with logistic regression

- We can get the same decision boundary with an infinite number of settings for $\mathbf{w}$.
- When the data are *separable* by $w_0 + \alpha \mathbf{w} \cdot \mathbf{x} = 0$, what's the best choice for α?

$$p(y = 1 \,|\, \mathbf{x}) \;=\; \sigma(w_0 + \alpha \mathbf{w} \cdot \mathbf{x}).$$

- With $\alpha \to \infty$, we have $p(y_i | \mathbf{x}; w_0, \alpha \mathbf{w}) \to 1$.

Overfitting with logistic regression

- We can get the same decision boundary with an infinite number of settings for $\mathbf{w}$.
- When the data are *separable* by $w_0 + \alpha \mathbf{w} \cdot \mathbf{x} = 0$, what's the best choice for α?

$$p(y = 1 \,|\, \mathbf{x}) \;=\; \sigma(w_0 + \alpha \mathbf{w} \cdot \mathbf{x}).$$

- With $\alpha \to \infty$, we have $p(y_i | \mathbf{x}; w_0, \alpha \mathbf{w}) \to 1$.
- With $\alpha = \infty$ there is a continuum of w_0 that reach perfect separation.
- When the data are not separable, similar effect is present but more subtle.

MAP estimation for logistic regression

- Intuition: we may have some belief about the value of $\mathbf{w}$ before seeing any data.
  - E.g., may prefer smaller values of $\|\mathbf{w}\|$ (ignore w_0)
    Recall our previous motivation for regularizing $\mathbf{w}$!
- A possible prior that captures that belief:

$$p(\mathbf{w}) \;=\; \mathcal{N}\left(\mathbf{w};\, \mathbf{0}, \sigma^2 \mathbf{I}\right).$$

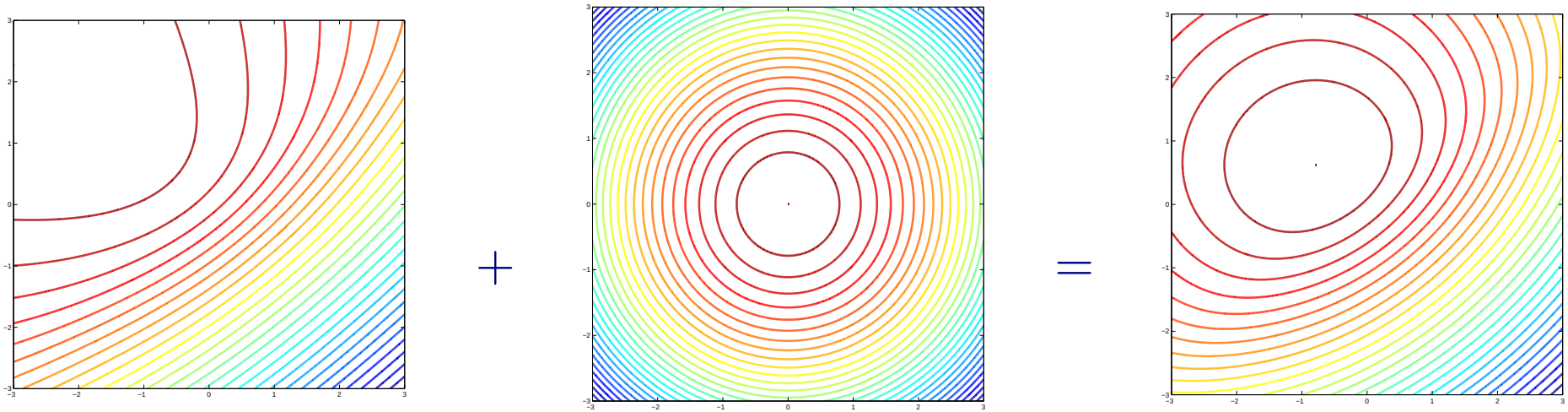
- Instead of $\log p(Y|X;\mathbf{w})$ the objective becomes log-posterior

$$\begin{aligned}\log p(Y|X,\mathbf{w};\sigma) \;&=\; \log p(Y|X,\mathbf{w}) + \log p(\mathbf{w};\sigma) \\ &= \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i, \mathbf{w}\right) \;-\; \frac{1}{2\sigma^2} \sum_{j=1}^{d} w_j^2 \;+ \text{const}(\mathbf{w}).\end{aligned}$$

- Setting σ^2 affects the penalty on $\|\mathbf{w}\|$ (cf. λ)

Log posterior surface

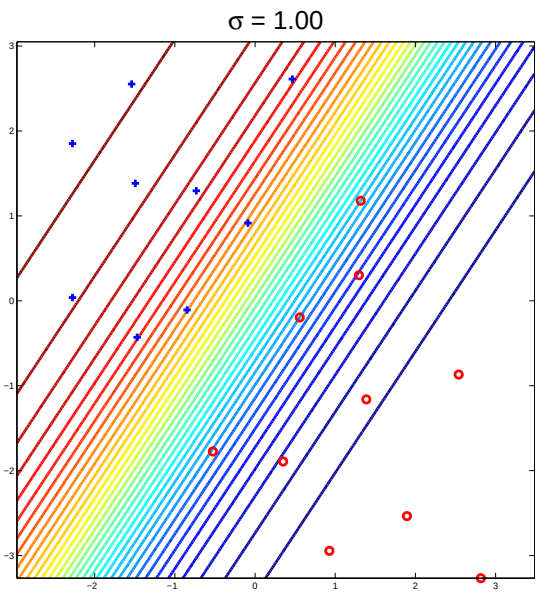
$\log p(Y|X, \mathbf{w})$ $\log p(\mathbf{w}; \sigma)$ $\log p(Y|X, \mathbf{w}; \sigma)$

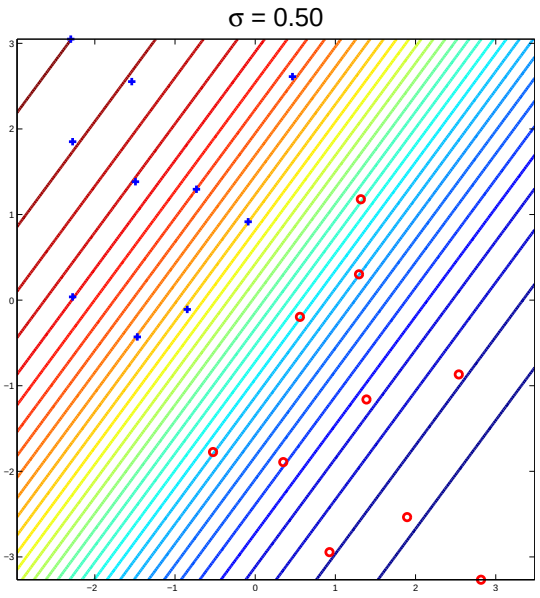


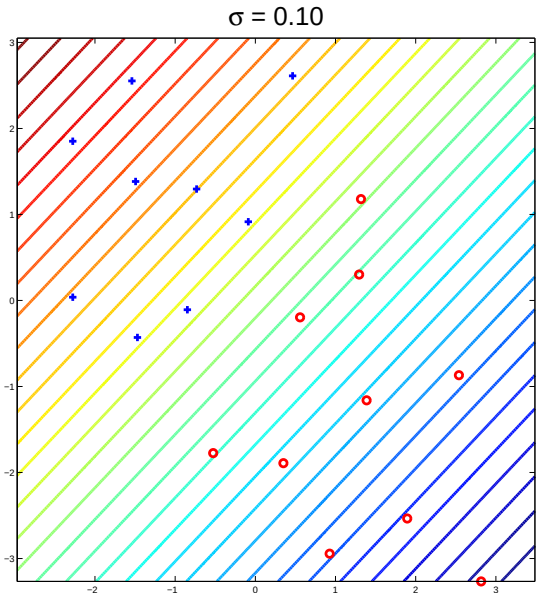
- This is our objective function, and we can find its peak by gradient ascent as before.
  - Need to modify the calculation of gradient and Hessian.

The effect of regularization: separable data

$$\log p(Y|X, \mathbf{w}; \sigma) = \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}\right) - \frac{1}{2\sigma^2}\|\mathbf{w}\|^2$$





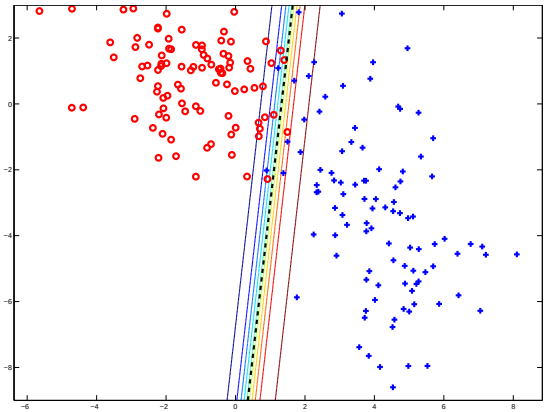


$\sigma^2 = 1$ $\sigma^2 = 0.5$ $\sigma^2 = 0.1$

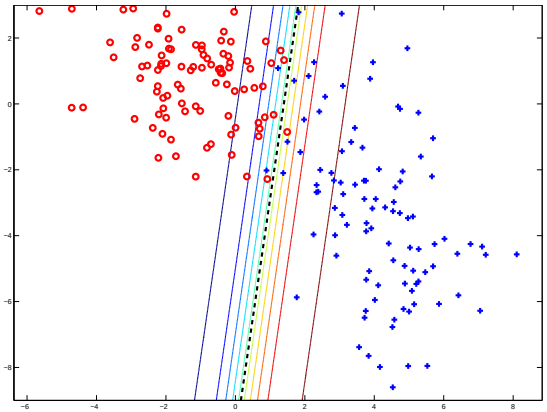
The effect of regularization

$$\log p(Y|X;\mathbf{w},\sigma) = \sum_{i=1}^{N} \log p\left(y_i \,|\, \mathbf{x}_i;\mathbf{w}\right) - \frac{1}{2\sigma^2}\|\mathbf{w}\|^2$$

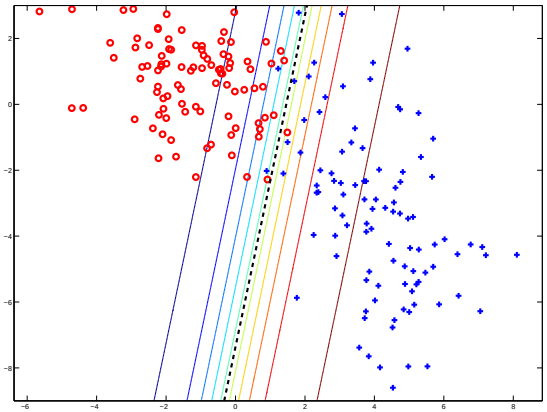
ML



$\sigma^2 = 1$



$\sigma^2 = 0.1$



$\sigma^2 = 0.01$