EN.600.475 Machine Learning

Classification

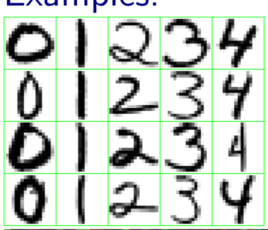
Raman Arora
Lecture 9
February 22, 2017

- Introduction to classification
- Logistic regression

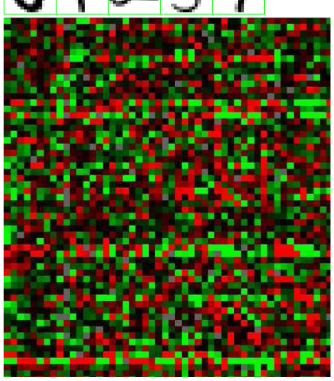
Slides credit: Greg Shakhnarovich

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?
- How do we decide on the label based on $f(\mathbf{x}; \hat{\mathbf{w}})$?

Intro to classification

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*:

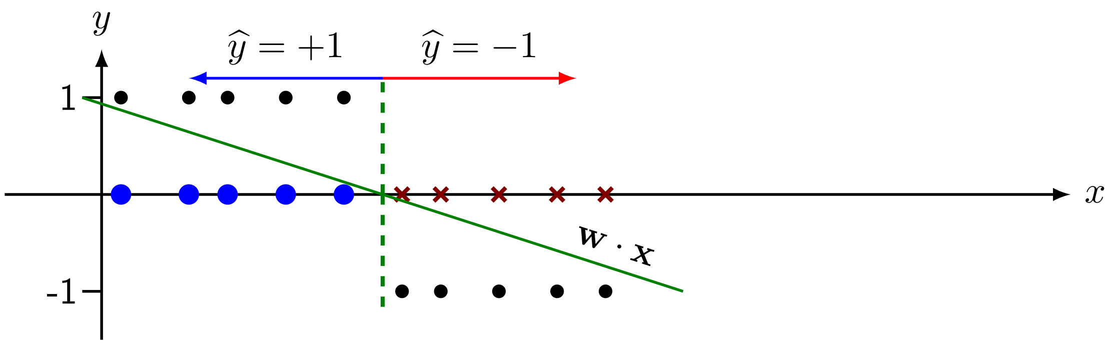
  decide on $\hat{y} = 1$ if $f(\mathbf{x}; \hat{\mathbf{w}}) \geq 0$, otherwise $\hat{y} = -1$.

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

- 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

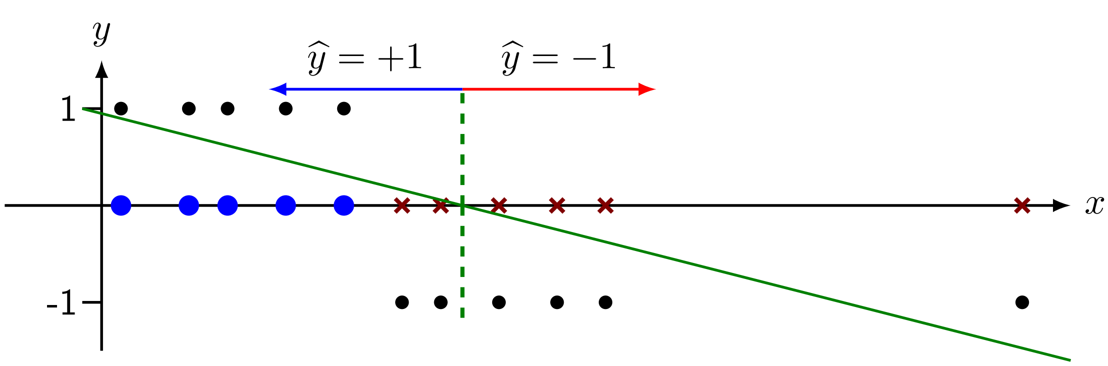
- A 1D example:



Intro to classification

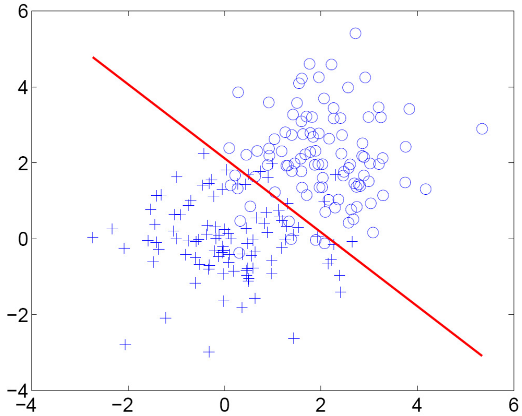
Classification as regression: example

- A 1D example:

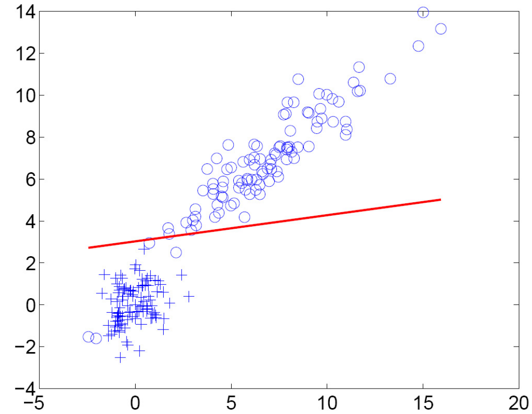


Classification as regression

- Same effect in 2D:

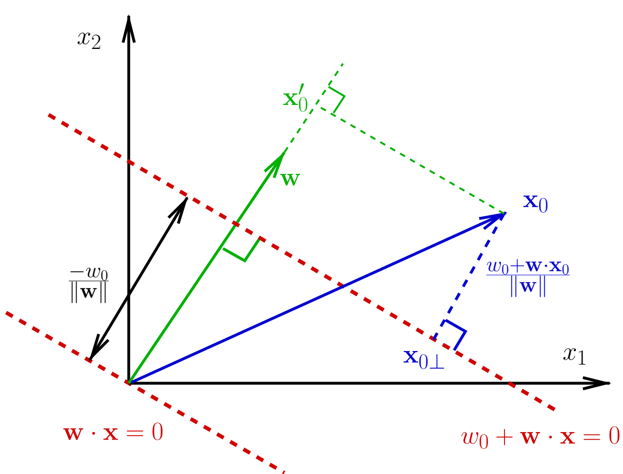


Seems to work well here



but not so well here

Geometry of projections



- $\mathbf{w} \cdot \mathbf{x} = 0$: 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: $\mathbf{x} \to (w_0 + \mathbf{w} \cdot \mathbf{x})/\|\mathbf{w}\|$.

Linear classifiers

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

- 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 $h(\mathbf{x})$:

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

- 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

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

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

$$
h(\mathbf{x}) = \underset{c}{\operatorname{argmax}}\, p\left(y = c \mid \mathbf{x}\right).
$$

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

Log-odds ratio

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

$$
\begin{aligned}
h(\mathbf{x}) = c^* \quad &\Leftrightarrow \quad \frac{p\left(y = c^* \mid \mathbf{x}\right)}{p\left(y = c \mid \mathbf{x}\right)} \geq 1 \quad \forall c \\
&\Leftrightarrow \quad \log \frac{p\left(y = c^* \mid \mathbf{x}\right)}{p\left(y = c \mid \mathbf{x}\right)} \geq 0 \quad \forall c
\end{aligned}
$$

- For the binary case,

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

The logistic model

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

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

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

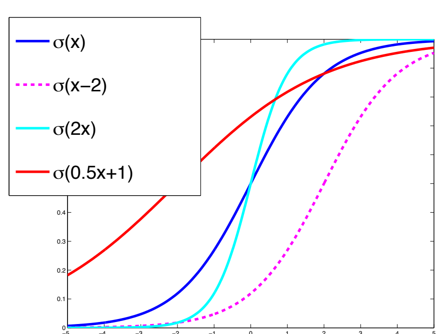
$$\begin{aligned}
\frac{p\,(y=1\,|\,\mathbf{x})}{1-p\,(y=1\,|\,\mathbf{x})} \;&=\; \exp(w_0 + \mathbf{w}\cdot\mathbf{x}) \;=\; 1\\
\Rightarrow \frac{1}{p\,(y=1\,|\,\mathbf{x})} \;&=\; 1+\exp(-w_0 - \mathbf{w}\cdot\mathbf{x}) \;=\; 2\\
\Rightarrow p\,(y=1\,|\,\mathbf{x}) \;&=\; \frac{1}{1+\exp(-w_0 -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

The logistic function

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

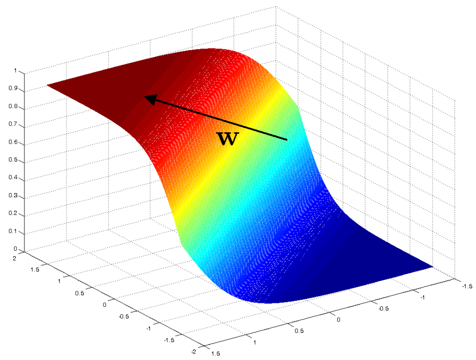
- The logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$:
  For any x, $0 \le \sigma(x) \le 1$;
  Monotonic, $\sigma(-\infty) = 0$, $\sigma(+\infty) = 1$

- $\sigma(0) = 1/2$. 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 $\mathbf{x} \in \mathbb{R}^d = [x_1 \ldots x_d]$?
- $\sigma(w_0 + \mathbf{w} \cdot \mathbf{x})$ is a scalar function of a scalar variable $w_0 + \mathbf{w} \cdot \mathbf{x}$.

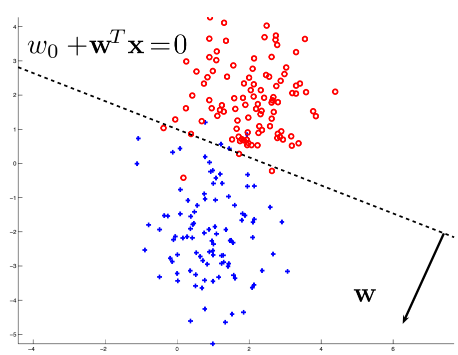


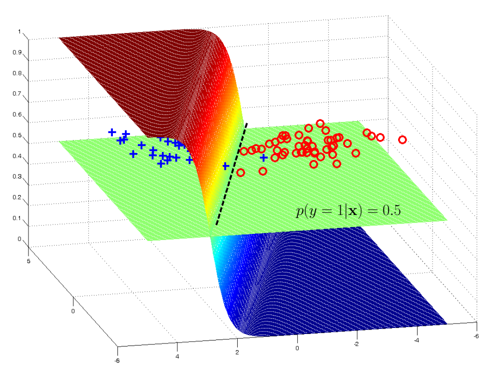
- the direction of $\mathbf{w}$ determines orientation;
- w_0 determines the location;
- $\|\mathbf{w}\|$ determines the slope.

Logistic regression: decision boundary

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

- 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.

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

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

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

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

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

- 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:

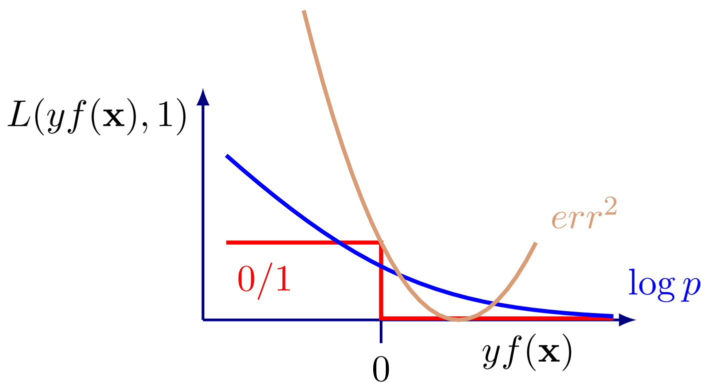
$$\mathbf{H} \; = \; \begin{bmatrix} \frac{\partial^2 \log p}{\partial w_0^2} & \frac{\partial^2 \log p}{\partial w_0 w_1} & \cdots & \frac{\partial^2 \log p}{\partial w_0 w_d} \\ \frac{\partial^2 \log p}{\partial w_0 w_1} & \frac{\partial^2 \log p}{\partial w_1^2} & \cdots & \frac{\partial^2 \log p}{\partial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

Surrogate loss

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

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

- 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):

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

- Example: quadratic logistic regression in 2D

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

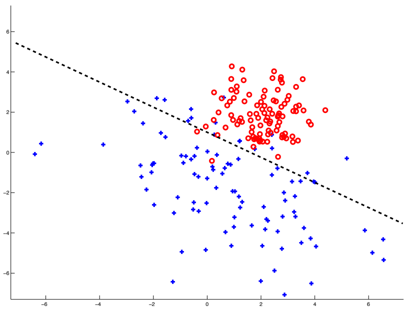
  - Decision boundary of this classifier:

$$w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 = 0,$$

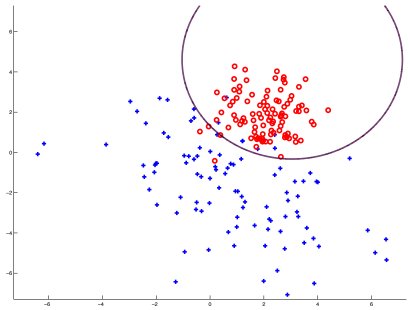
    i.e. it's a quadratic decision boundary.

Logistic regression: 2D example

Linear



Quadratic



We can also include $x_1 x_2$: