exp[log(p/(1-p))(math=55) – log(p/(1-p))(math = 54)] =
odds(math=55)/odds(math=54) = exp(.1563404) = 1.1692241

0.156 is the difference of linear part

Lecture 7: Logistic Regression

TTIC 31020: Introduction to Machine Learning

Instructor: Kevin Gimpel

TTI-Chicago

October 22, 2019

Review: geometry of regularization

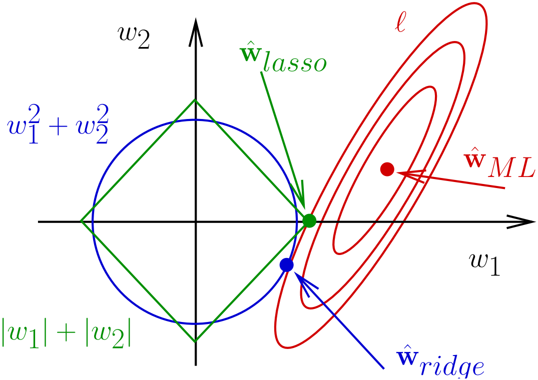
Can write unconstrained optimization problem

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}\cdot\mathbf{x}_i)^2 + \lambda \sum_{j=1}^{m} |w_j|^p$$

as an equivalent constrained problem:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}\cdot\mathbf{x}_i)^2$$

$$\text{subject to } \sum_{j=1}^{m} |w_j|^p \le t$$



$p = 1$ may lead to sparsity, $p = 2$ generally won't

Roadmap

So far:

- General principles: empirical loss, (expected) risk, training/test data
- Formulating learning as optimization
- Generalized linear least squares regression
- Gradient descent as a learning algorithm

Today:

- Linear models for classification
- Stochastic gradient descent

Then:

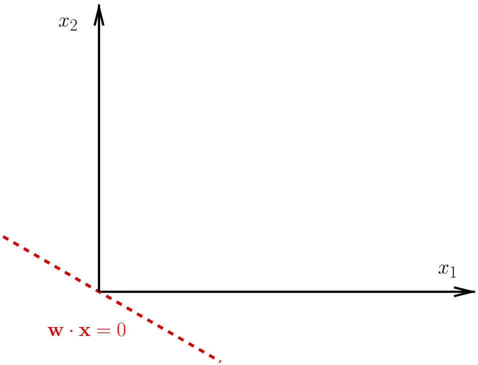
- Support vector machines; kernels

Geometry of projections

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

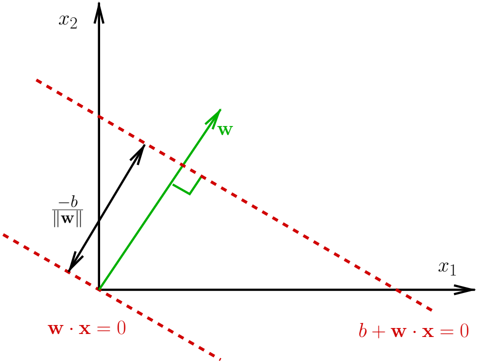
Geometry of projections

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



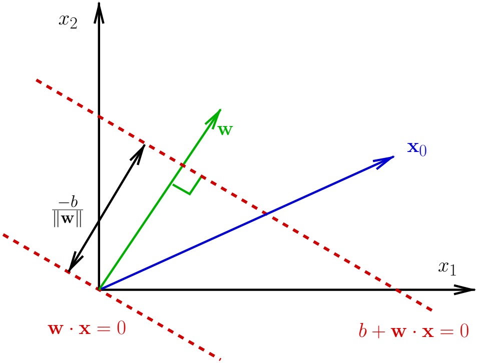
Geometry of projections

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



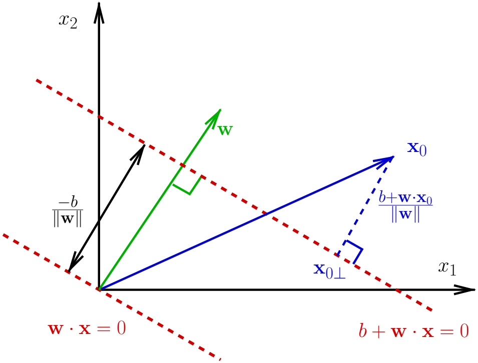
Geometry of projections

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



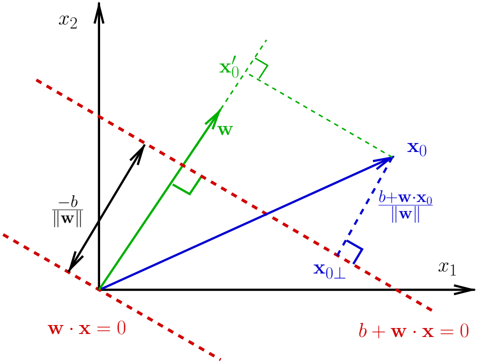
Geometry of projections

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



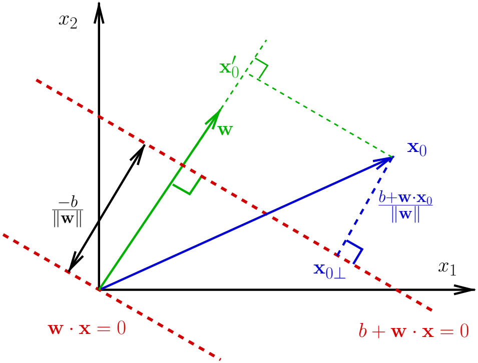
Geometry of projections

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



Geometry of projections

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



- Set up a new 1D coordinate system defined by projection of $\mathbf{x}$ onto the vector $\mathbf{w}$:
  $\mathbf{x} \;\rightarrow\; (b + \mathbf{w} \cdot \mathbf{x})/\|\mathbf{w}\|$

Linear classifiers

$$\widehat{y} = h(\mathbf{x}) \;=\; \operatorname{sign}(b + \mathbf{w} \cdot \mathbf{x})$$

- Classifying using a linear decision boundary effectively reduces the data dimension to 1
- Need to find $\mathbf{w}$ (direction) and b (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

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

Risk of a classifier

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

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

Risk of a classifier

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

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

Risk of a classifier

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

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

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

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

Conditional risk of a classifier

$$R(h \mid \mathbf{x}) \;=\; \sum_{c=1}^{C} \ell(h(\mathbf{x}), c) p\,(y = c \mid \mathbf{x})$$

Conditional risk of a classifier

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

Conditional risk of a classifier

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

Conditional risk of a classifier

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

- 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 $h(\mathbf{x}) = \operatorname{argmax}_c p\,(y = c \,|\, \mathbf{x})$ is equivalent to

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

Log-odds ratio

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

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

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

- For the binary case,

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

The logistic model

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

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

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

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

The logistic model

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

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

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

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

The logistic model

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

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

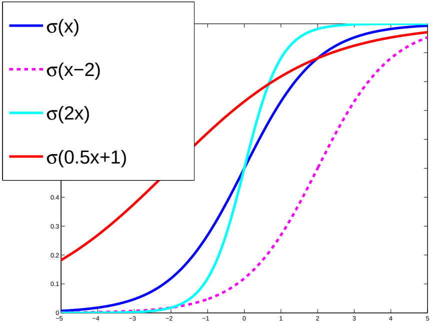
- Since $p\,(y=1\,|\,\mathbf{x}) = 1 - p\,(y=0\,|\,\mathbf{x})$, we have (after exponentiating):

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

The logistic function

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

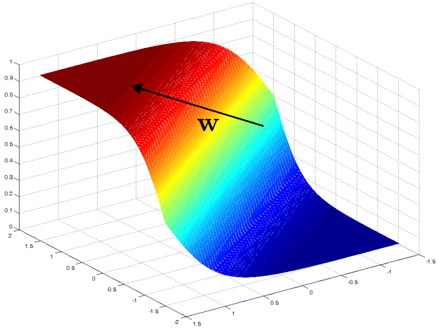
- The logistic (sigmoid) function: $\sigma(x) = \frac{1}{1+e^{-x}}$
  For any x, $0 \leq \sigma(x) \leq 1$
  Monotonic, $\sigma(-\infty) = 0$, $\sigma(+\infty) = 1$



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

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

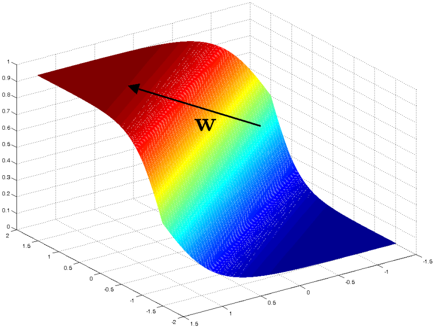
- What if $\mathbf{x} \in \mathbb{R}^d \ = [x_1 \ldots \ x_d]$?
- $\sigma(b + \mathbf{w} \cdot \mathbf{x})$ is a scalar function of a scalar variable $b + \mathbf{w} \cdot \mathbf{x}$.



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

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

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

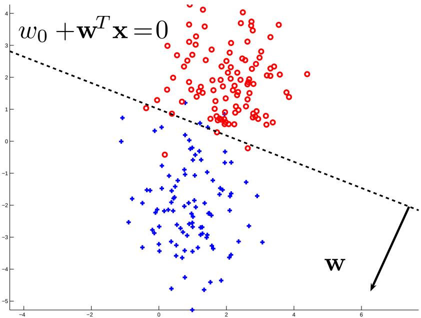


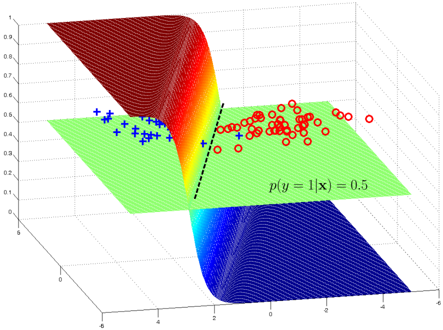
- the direction of $\mathbf{w}$ determines orientation
- b determines the location
- $\|\mathbf{w}\|$ determines the slope

Logistic regression: decision boundary

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

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





Complexity of logistic regression

- We can choose a set of features (basis functions):

$$p\,(y = 1\,|\,\mathbf{x}) \;=\; \sigma\,(\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}))$$

Complexity of logistic regression

- We can choose a set of features (basis functions):

$$p\left(y=1 \,|\, \mathbf{x}\right) \;=\; \sigma\left(\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})\right)$$

- Example: quadratic logistic regression in 2D

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

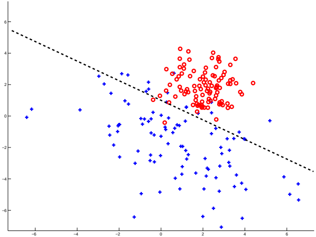
- Decision boundary of this classifier:

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

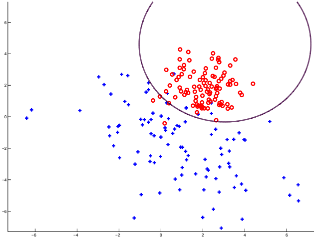
  i.e., it's a quadratic decision boundary.

Logistic regression: 2D example

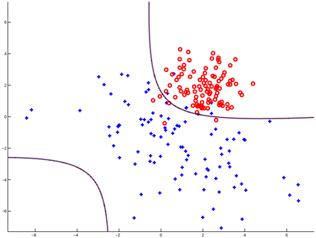
Linear



Quadratic



We can also include $x_1 x_2$:



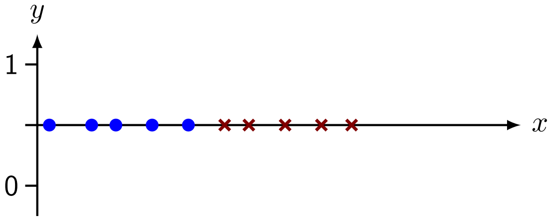
Likelihood under the logistic model

- Least squares regression: minimize squared residuals vs. observed y
- Logistic regression: directly maximize *probability* of observed $y|\mathbf{x}$

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

$$\log p(\mathbf{y} \mid \mathbf{X}; \mathbf{w}, b) \;= \sum_{i=1}^{n} y_i \log \sigma(b + \mathbf{w} \cdot \mathbf{x}_i) + (1 - y_i) \log\left(1 - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)\right)$$

- We can treat $y_i - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)$ as the **prediction error** of the model on $\mathbf{x}_i, y_i$



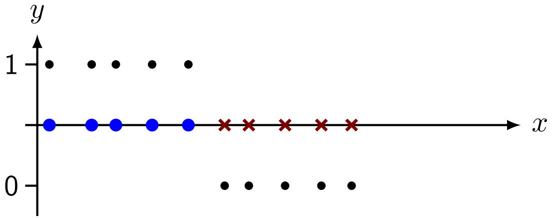
Likelihood under the logistic model

- Least squares regression: minimize squared residuals vs. observed y
- Logistic regression: directly maximize *probability* of observed $y|\mathbf{x}$

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

$$
\log p(\mathbf{y} \mid \mathbf{X}; \mathbf{w}, b) = \sum_{i=1}^{n} y_i \log \sigma(b + \mathbf{w} \cdot \mathbf{x}_i) + (1 - y_i) \log\left(1 - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)\right)
$$

- We can treat $y_i - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)$ as the **prediction error** of the model on $\mathbf{x}_i, y_i$



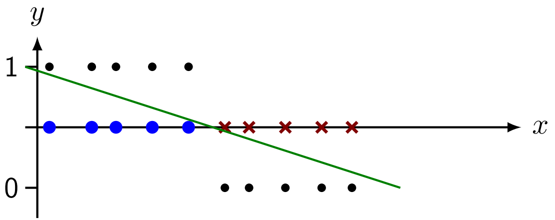
Likelihood under the logistic model

- Least squares regression: minimize squared residuals vs. observed y
- Logistic regression: directly maximize *probability* of observed $y|\mathbf{x}$

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

$$\log p(\mathbf{y} \mid \mathbf{X}; \mathbf{w}, b) = \sum_{i=1}^{n} y_i \log \sigma(b + \mathbf{w} \cdot \mathbf{x}_i) + (1 - y_i) \log \left(1 - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)\right)$$

- We can treat $y_i - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)$ as the **prediction error** of the model on $\mathbf{x}_i, y_i$



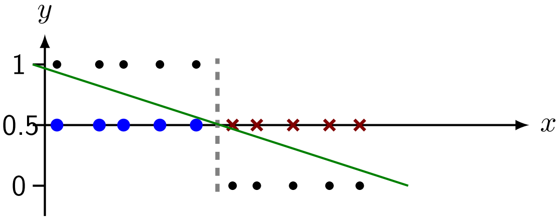
Likelihood under the logistic model

- Least squares regression: minimize squared residuals vs. observed y
- Logistic regression: directly maximize *probability* of observed $y|\mathbf{x}$

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

$$\log p(\mathbf{y} \mid \mathbf{X}; \mathbf{w}, b) = \sum_{i=1}^{n} y_i \log \sigma(b + \mathbf{w} \cdot \mathbf{x}_i) + (1 - y_i) \log\left(1 - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)\right)$$

- We can treat $y_i - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)$ as the **prediction error** of the model on $\mathbf{x}_i, y_i$



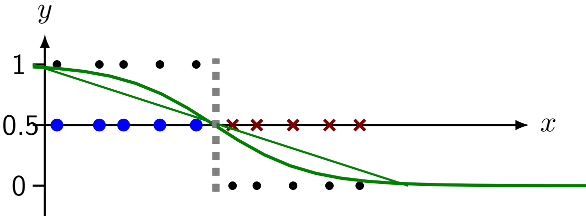
Likelihood under the logistic model

- Least squares regression: minimize squared residuals vs. observed y
- Logistic regression: directly maximize *probability* of observed $y|\mathbf{x}$

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

$$\log p(\mathbf{y} \mid \mathbf{X}; \mathbf{w}, b) = \sum_{i=1}^{n} y_i \log \sigma(b + \mathbf{w} \cdot \mathbf{x}_i) + (1 - y_i) \log\left(1 - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)\right)$$

- We can treat $y_i - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)$ as the **prediction error** of the model on $\mathbf{x}_i, y_i$



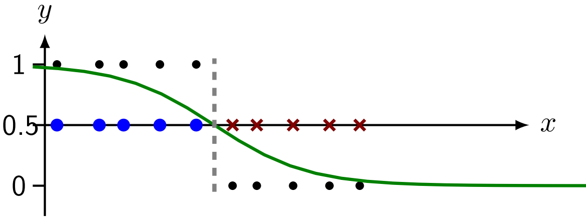
Likelihood under the logistic model

- Least squares regression: minimize squared residuals vs. observed y
- Logistic regression: directly maximize *probability* of observed $y|\mathbf{x}$

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

$$\log p(\mathbf{y} \mid \mathbf{X}; \mathbf{w}, b) = \sum_{i=1}^{n} y_i \log \sigma(b + \mathbf{w} \cdot \mathbf{x}_i) + (1 - y_i) \log\left(1 - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)\right)$$

- We can treat $y_i - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)$ as the **prediction error** of the model on $\mathbf{x}_i, y_i$



Likelihood under the logistic model

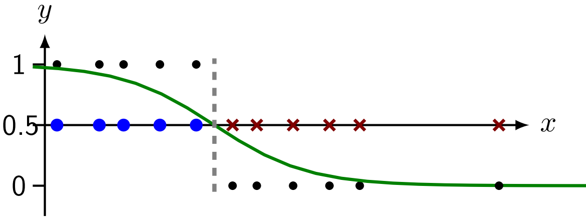
- Least squares regression: minimize squared residuals vs. observed y
- Logistic regression: directly maximize *probability* of observed $y|\mathbf{x}$

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

$$= \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)^{y_i} \left(1 - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)\right)^{1-y_i}$$

$$\log p(\mathbf{y} \mid \mathbf{X}; \mathbf{w}, b) = \sum_{i=1}^{n} y_i \log \sigma(b + \mathbf{w} \cdot \mathbf{x}_i) + (1 - y_i) \log\left(1 - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)\right)$$

- We can treat $y_i - \sigma(b + \mathbf{w} \cdot \mathbf{x}_i)$ as the **prediction error** of the model on $\mathbf{x}_i, y_i$



Gradient descent

- Generally, we predict y from $\boldsymbol{\phi}(\mathbf{x})$
- We can cycle through the examples, accumulating the gradient, and then apply the accumulated value to form an update
- Initialize $\mathbf{w}^{(0)} = \mathbf{0}$
- For t until convergence: calculate gradient,

$$\nabla_{\mathbf{w}}^{(t)} \log p\left(y_i \,|\, \mathbf{x}_i; \mathbf{w}^{(t)}\right) \;=\; \left[y_i - \sigma(\mathbf{w}^{(t)} \cdot \boldsymbol{\phi}(\mathbf{x}_i))\right] \boldsymbol{\phi}(\mathbf{x}_i)$$

  update model

$$\mathbf{w}^{(t+1)} \;=\; \mathbf{w}^{(t)} \;+\; \eta \nabla_{\mathbf{w}}^{(t)}$$

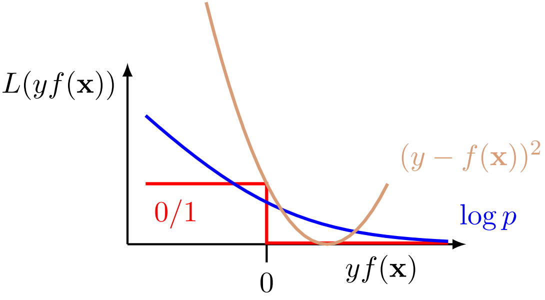
- Remember: need to choose η carefully:
- Too small $\Rightarrow$ slow convergence
- Too large: $\Rightarrow$ overshoot and oscillate

Surrogate loss

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

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

- This is a **surrogate** loss; we use it because it's not computationally feasible to optimize 0/1 loss directly



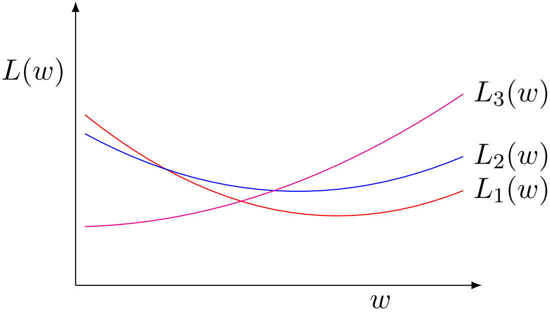
- Can redefine loss in terms of **margin** $yf(\mathbf{x})$ when $y \in \{\pm 1\}$

Stochastic gradient descent: intuition

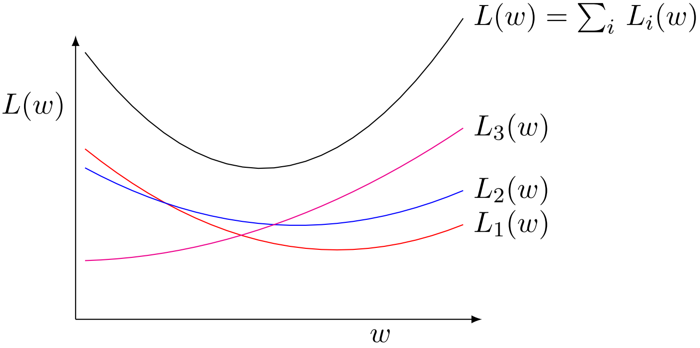
- Computing gradient on all n examples is computationally expensive and may be unnecessary
- Many data points provide similar information
- Idea: present examples one at a time, and pretend that the gradient on the entire set is the same as gradient on one example
- Formally: estimate gradient of the loss L

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \mathbf{w}}L(y_i,\mathbf{x}_i;\mathbf{w}) \approx \frac{\partial}{\partial \mathbf{w}}L(y_t,\mathbf{x}_t;\mathbf{w})$$

Stochastic gradient descent: intuition

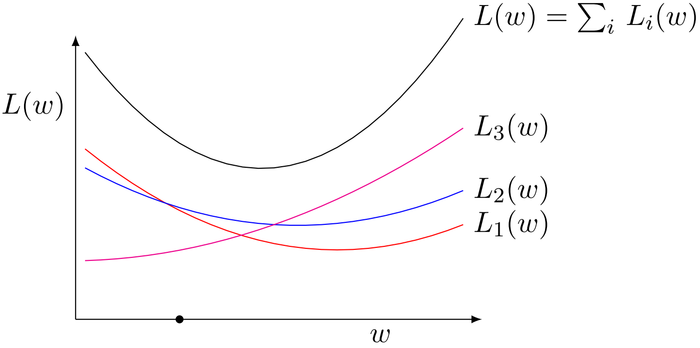


Stochastic gradient descent: intuition



- Objective: $\min_w L(w) = \min_w \sum_{i=1}^{n} L_i(w)$

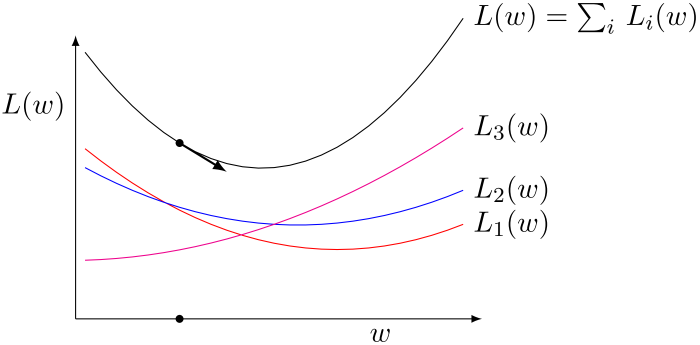
Stochastic gradient descent: intuition



- Objective: $\min_w L(w) = \min_w \sum_{i=1}^{n} L_i(w)$
- Stochastic approximation: given an i, estimate

$$\frac{1}{n}\nabla L(w) \approx \nabla L_i(w)$$

Stochastic gradient descent: intuition

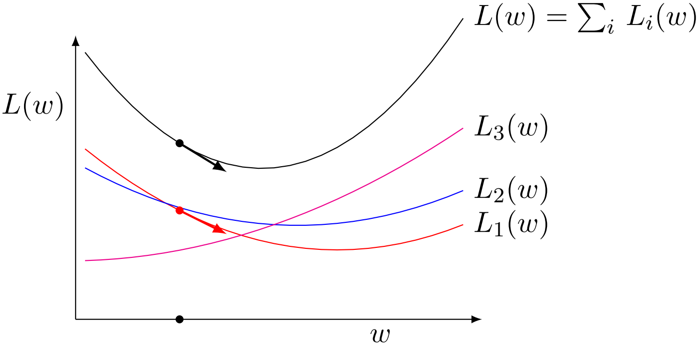


- Objective: $\min_w L(w) = \min_w \sum_{i=1}^n L_i(w)$
- Stochastic approximation: given an i, estimate

$$\frac{1}{n} \nabla L(w) \approx \nabla L_i(w)$$

may help in penalize since it never reach the global maximum

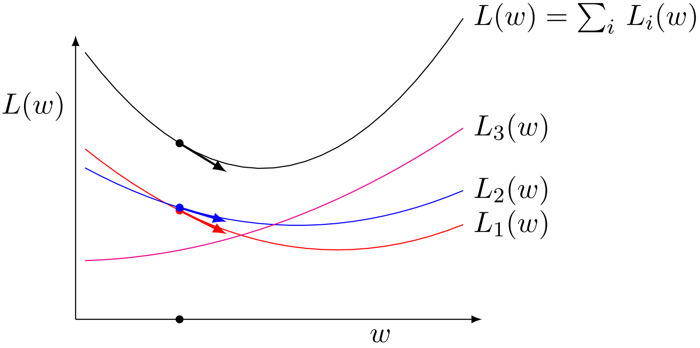
Stochastic gradient descent: intuition



- Objective: $\min_w L(w) = \min_w \sum_{i=1}^n L_i(w)$
- Stochastic approximation: given an i, estimate

$$\frac{1}{n}\nabla L(w) \approx \nabla L_i(w)$$

Stochastic gradient descent: intuition

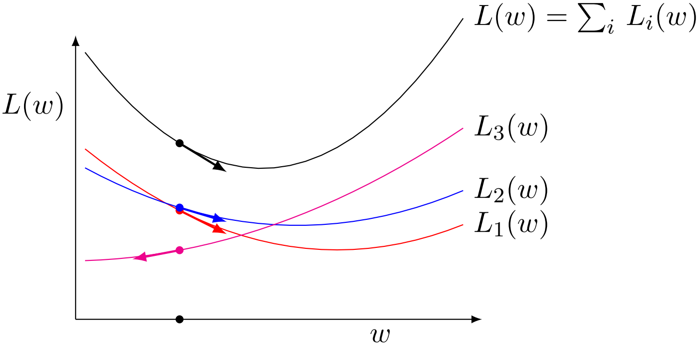


- Objective: $\min_w L(w) = \min_w \sum_{i=1}^{n} L_i(w)$
- Stochastic approximation: given an i, estimate

$$\frac{1}{n}\nabla L(w) \approx \nabla L_i(w)$$

- Could be a noisy estimate

Stochastic gradient descent: intuition



- Objective: $\min_w L(w) = \min_w \sum_{i=1}^{n} L_i(w)$
- Stochastic approximation: given an i, estimate

$$\frac{1}{n}\nabla L(w) \approx \nabla L_i(w)$$

- Could be a noisy estimate