Lecture 26 – Logistic Regression and Maximum Likelihood Estimation (continued)

DSC 40A, Fall 2022 @ UC San Diego
Dr. Truong Son Hy, with help from **many others**
Some materials are taken from Prof. Greg Shakhnarovich's ML course at TTIC.

Announcements

- The final is coming!
- There will be a review session.

Agenda

- Logistic Regression.
- Maximum Likelihood Estimation.

Logistic Regression

Linear classifier

- Hypothesis:
$$\hat{y} = h(\vec{x}) = \mathrm{sign}(\vec{x} \cdot \vec{w} + w_0)$$
- Classifying using a linear decision boundary effectively reduces the data dimension to 1.
- We need to find the direction $\vec{w}$ and location w_0 of the boundary.
- We want to minimize the expected **zero/one** loss for classifier $h : X \to Y$, which for $(\vec{x}, y)$ is:

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

Empirical Risk Minimization

- The risk (expected loss) of a C-way classifier $h(\vec{x})$ (i.e. C is the number of classes):

$$R(h) = E_{p(\vec{x},y)}[L(h(\vec{x}), y)],$$

  where E denotes the expectation and $p(\vec{x}, y)$ denotes the joint probability distribution of our data $(\vec{x}, y)$. Our data is considered as samples drawn from p.

- We can write the risk in intergral form:

$$R(h) = \int_{\vec{x}} \sum_{c=1}^{C} L(h(\vec{x}), c) p(\vec{x}, y = c) d\vec{x}$$

Empirical Risk Minimization

- We can further write the risk as:

$$R(h) = \int_{\vec{x}} \left[\sum_{c=1}^{C} L(h(\vec{x}), c) p(y = c | \vec{x}) \right] p(\vec{x}) d\vec{x}$$

- Clearly, it is enough to minimize the **conditional risk** for any $\vec{x}$:

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

Conditional risk of a classifier

- Conditional risk:

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

- We can factorize this risk as:

$$R(h|\vec{x}) = 0 \cdot p(y = h(\vec{x})|\vec{x}) + 1 \cdot \sum_{c \neq h(\vec{x})} p(y = c|\vec{x}),$$

$$\Leftrightarrow R(h|\vec{x}) = \sum_{c \neq h(\vec{x})} p(y = c|\vec{x}) = 1 - p(y = h(\vec{x})|\vec{x})$$

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

$$h(\vec{x}) = \mathrm{argmax}_c p(y = c|\vec{x})$$

Log-odds ratio

- Optimal rule $h(\vec{x}) = \text{argmax}_c p(y = c|\vec{x})$ is equivalent to:

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

  that is equivalent to:

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

- For the binary case:

$$h(\vec{x}) = 1 \Leftrightarrow \log \frac{p(y = 1|\vec{x})}{p(y = 0|\vec{x})} \geq 0.$$

The logistic model

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

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

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

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

The logistic model

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

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

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

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

$$\Rightarrow \frac{1}{p(y = 1|\vec{x})} = 1 + \exp(-\vec{x} \cdot \vec{w} - w_0) = 2$$

The logistic model

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

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

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

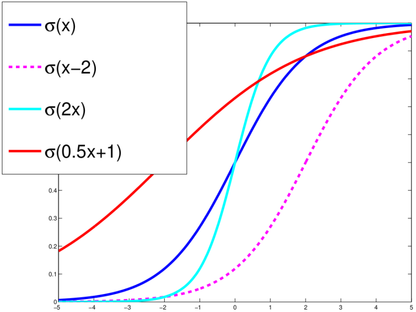
$$\frac{p(y=1|\vec{x})}{1-p(y=1|\vec{x})} = \exp(\vec{x} \cdot \vec{w} + w_0) = 1$$

$$\Rightarrow \frac{1}{p(y=1|\vec{x})} = 1 + \exp(-\vec{x} \cdot \vec{w} - w_0) = 2$$

$$\Rightarrow p(y=1|\vec{x}) = \frac{1}{1+\exp(-\vec{x} \cdot \vec{w} - w_0)} = \frac{1}{2}$$

The logistic function

- The logistic / sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$.
  For any $x \in R$: $0 \leq \sigma(x) \leq 1$.
  Monotonic: $\lim_{x \to -\infty} \sigma(x) = 0$ and $\lim_{x \to +\infty} \sigma(x) = 1$.

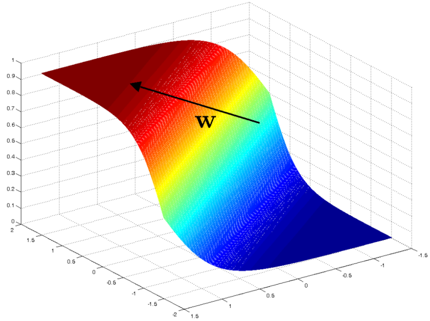


- We have:

$$p(y = 1 | \vec{x}) = \frac{1}{1 + \exp(-\vec{x} \cdot \vec{w} - w_0)} = \sigma(h(\vec{x}))$$

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

- For $\vec{x} \in \mathbb{R}^d$, $\sigma(\vec{w} \cdot \vec{x} + w_0)$ is a scalar function of a scalar variable $\vec{w} \cdot \vec{x} + w_0$.

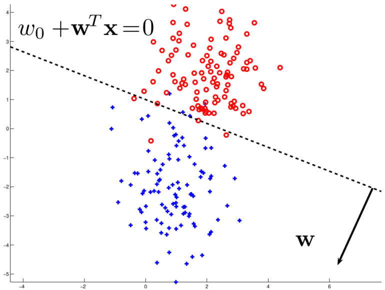


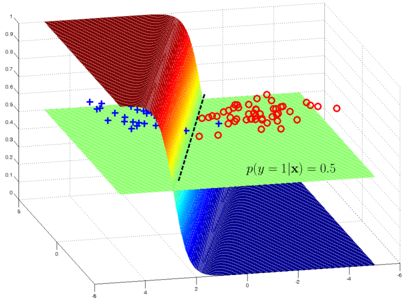
- The direction of $\vec{w}$ determines the orientation, w_0 determines the location, and $\|\vec{w}\|$ determines the slope.

Decision boundary of Logistic Regression

With linear logistic model, we get a linear decision boundary:

$$p(y = 1|\vec{x}) = \sigma(\vec{w} \cdot \vec{x} + w_0) = \frac{1}{2} \Leftrightarrow \vec{w} \cdot \vec{x} + w_0 = 0$$





Maximum Likelihood Estimation

Likelihood under the logistic model

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

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

We can write it compactly as:

$$p(y_i|\vec{x}_i;\vec{w},w_0) = \sigma(\vec{w}\cdot\vec{x}_i + w_0)^{y_i} \cdot (1 - \sigma(\vec{w}\cdot\vec{x}_i + w_0))^{1-y_i}.$$

Likelihood under the logistic model

Suppose we are given a dataset $D = \{(\vec{x}_i, y_i)\}_{i=1}^N$ of N samples. The **likelihood** of $\vec{w}$ and w_0 on this data is defined as:

$$p(Y|X; \vec{w}, w_0) = \prod_{i=1}^{N} p(y_i|\vec{x}_i; \vec{w}, w_0)$$

The **log-likelihood** is then:

$$\log p(Y|X; \vec{w}, w_0) =$$

Likelihood under the logistic model

Suppose we are given a dataset $D = \{(\vec{x}_i, y_i)\}_{i=1}^{N}$ of N samples. The **likelihood** of $\vec{w}$ and w_0 on this data is defined as:

$$p(Y|X;\vec{w},w_0) = \prod_{i=1}^{N} p(y_i|\vec{x}_i;\vec{w},w_0)$$

The **log-likelihood** is then:

$$\log p(Y|X;\vec{w},w_0) = \sum_{i=1}^{N} \log p(y_i|\vec{x}_i;\vec{w},w_0),$$

that is equal to:

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

Maximum Likelihood Solution

We want to find $\vec{w}$ and w_0 that **maximizes** the log-likelihood:

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

Maximum Likelihood Solution

We want to find $\vec{w}$ and w_0 that **maximizes** the log-likelihood:

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

We find the derivatives:

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

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

We can treat $y_i - p(y_i|\vec{x}_i) = y_i - \sigma(\vec{w}\cdot\vec{x}_i + w_0)$ as the **prediction error** of the model on $\vec{x}_i, y_i$.

Derivatives for $\ln$ and σ

- **Logarithm:**

$$\frac{d}{dx}\ln(x) = \frac{1}{x}, \qquad \frac{d}{dx}\log_a(x) = \frac{1}{x\ln(a)}$$

Derivatives for $\ln$ and σ

- **Logarithm:**

$$\frac{d}{dx}\ln(x) = \frac{1}{x}, \qquad \frac{d}{dx}\log_a(x) = \frac{1}{x\ln(a)}$$

- **Sigmoid:**

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}\frac{1}{1+e^{-x}} = \frac{d}{dx}(1+e^{-x})^{-1} = -(1+e^{-x})^{-2}\frac{d}{dx}(1+e^{-x})$$

Derivatives for $\ln$ and σ

- **Logarithm:**

$$\frac{d}{dx}\ln(x) = \frac{1}{x}, \qquad \frac{d}{dx}\log_a(x) = \frac{1}{x\ln(a)}$$

- **Sigmoid:**

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}\frac{1}{1+e^{-x}} = \frac{d}{dx}(1+e^{-x})^{-1} = -(1+e^{-x})^{-2}\frac{d}{dx}(1+e^{-x})$$

$$\Leftrightarrow \frac{d}{dx}\sigma(x) = (1+e^{-x})^{-2}e^{-x} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}}$$

$$\Leftrightarrow \frac{d}{dx}\sigma(x) = \sigma(x) \cdot \frac{e^{-x}+1-1}{1+e^{-x}} = \sigma(x) \cdot [1-\sigma(x)]$$

Partial derivatives for $\log \sigma(\vec{w} \cdot \vec{x}_i + w_0)$

For w_0:

$$\frac{\partial}{\partial w_0} \log \sigma(\vec{w} \cdot \vec{x}_i + w_0) = \frac{1}{\sigma(\vec{w} \cdot \vec{x}_i + w_0)} \frac{\partial}{\partial w_0} \sigma(\vec{w} \cdot \vec{x}_i + w_0)$$

Partial derivatives for $\log \sigma(\vec{w} \cdot \vec{x}_i + w_0)$

For w_0:

$$\frac{\partial}{\partial w_0} \log \sigma(\vec{w} \cdot \vec{x}_i + w_0) = \frac{1}{\sigma(\vec{w} \cdot \vec{x}_i + w_0)} \frac{\partial}{\partial w_0} \sigma(\vec{w} \cdot \vec{x}_i + w_0)$$

$$= \frac{1}{\sigma(\vec{w} \cdot \vec{x}_i + w_0)} \sigma(\vec{w} \cdot \vec{x}_i + w_0)[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)] \frac{\partial}{\partial w_0} (\vec{w} \cdot \vec{x}_i + w_0)$$

Partial derivatives for $\log \sigma(\vec{w} \cdot \vec{x}_i + w_0)$

For w_0:

$$\frac{\partial}{\partial w_0} \log \sigma(\vec{w} \cdot \vec{x}_i + w_0) = \frac{1}{\sigma(\vec{w} \cdot \vec{x}_i + w_0)} \frac{\partial}{\partial w_0} \sigma(\vec{w} \cdot \vec{x}_i + w_0)$$

$$= \frac{1}{\sigma(\vec{w} \cdot \vec{x}_i + w_0)} \sigma(\vec{w} \cdot \vec{x}_i + w_0)[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)] \frac{\partial}{\partial w_0} (\vec{w} \cdot \vec{x}_i + w_0)$$

$$= 1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)$$

For w_j:

$$\frac{\partial}{\partial w_j} \log \sigma(\vec{w} \cdot \vec{x}_i + w_0) = \frac{1}{\sigma(\vec{w} \cdot \vec{x}_i + w_0)} \frac{\partial}{\partial w_j} \sigma(\vec{w} \cdot \vec{x}_i + w_0)$$

$$= \frac{1}{\sigma(\vec{w} \cdot \vec{x}_i + w_0)} \sigma(\vec{w} \cdot \vec{x}_i + w_0)[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)] \frac{\partial}{\partial w_j} (\vec{w} \cdot \vec{x}_i + w_0)$$

$$= [1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)] x_{ij}$$

Partial derivatives for $\log[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)]$

For w_0:

$$\frac{\partial}{\partial w_0} \log[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)] = \frac{1}{1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)} \frac{\partial}{\partial w_0}[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)]$$

$$= \frac{-1}{1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)} \sigma(\vec{w} \cdot \vec{x}_i + w_0)[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)] \frac{\partial}{\partial w_0}(\vec{w} \cdot \vec{x}_i + w_0)$$

$$= -\sigma(\vec{w} \cdot \vec{x}_i + w_0)$$

Partial derivatives for $\log[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)]$

For w_0:

$$\frac{\partial}{\partial w_0} \log[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)] = \frac{1}{1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)} \frac{\partial}{\partial w_0}[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)]$$

$$= \frac{-1}{1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)} \sigma(\vec{w} \cdot \vec{x}_i + w_0)[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)] \frac{\partial}{\partial w_0}(\vec{w} \cdot \vec{x}_i + w_0)$$

$$= -\sigma(\vec{w} \cdot \vec{x}_i + w_0)$$

For w_j:

$$\frac{\partial}{\partial w_j} \log[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)] = \frac{1}{1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)} \frac{\partial}{\partial w_j}[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)]$$

$$= \frac{-1}{1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)} \sigma(\vec{w} \cdot \vec{x}_i + w_0)[1 - \sigma(\vec{w} \cdot \vec{x}_i + w_0)] \frac{\partial}{\partial w_j}(\vec{w} \cdot \vec{x}_i + w_0)$$

$$= -\sigma(\vec{w} \cdot \vec{x}_i + w_0) x_{ij}$$

Partial derivatives for $\log p(Y|X;\vec{w},w_0)$

Log-likelihood:

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

For w_0:

$$\frac{\partial}{\partial w_0}\log p(Y|X;\vec{w},w_0) = \sum_{i=1}^{N} y_i[1-\sigma(\vec{w}\cdot\vec{x}_i+w_0)]+(1-y_i)[-\sigma(\vec{w}\cdot\vec{x}_i+w_0)]$$

$$= \sum_{i=1}^{N} y_i - y_i\sigma(\vec{w}\cdot\vec{x}_i + w_0) - \sigma(\vec{w}\cdot\vec{x}_i + w_0) + y_i\sigma(\vec{w}\cdot\vec{x}_i + w_0)$$

$$= \sum_{i=1}^{N}[y_i - \sigma(\vec{w}\cdot\vec{x}_i + w_0)]$$

Partial derivatives for $\log p(Y|X;\vec{w},w_0)$

Log-likelihood:

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

For w_j:

$$\frac{\partial}{\partial w_j}\log p(Y|X;\vec{w},w_0) = \sum_{i=1}^{N} y_i[1-\sigma(\vec{w}\cdot\vec{x}_i+w_0)]x_{ij}+(1-y_i)[-\sigma(\vec{w}\cdot\vec{x}_i+w_0)]x_{ij}$$

$$= \sum_{i=1}^{N}[y_i - y_i\sigma(\vec{w}\cdot\vec{x}_i + w_0) - \sigma(\vec{w}\cdot\vec{x}_i + w_0) + y_i\sigma(\vec{w}\cdot\vec{x}_i + w_0)]x_{ij}$$

$$= \sum_{i=1}^{N}[y_i - \sigma(\vec{w}\cdot\vec{x}_i + w_0)]x_{ij}$$

Gradient ascent for MLE

Thus, we get the gradients as follows:

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

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

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

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \alpha \cdot \frac{\partial}{\partial w_0} \log p(Y|X;\vec{w}, w_0)$$

$$\vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} + \alpha \cdot \frac{\partial}{\partial \vec{w}} \log p(Y|X;\vec{w}, w_0)\vec{x}$$

where α is the learning rate.

Gradient ascent for MLE

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

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

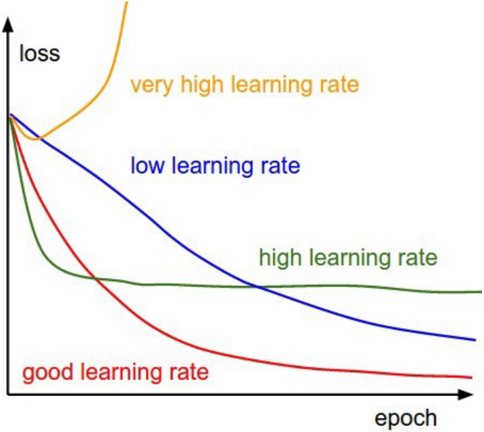
- This is a *surrogate* loss: we work with it since it is not computationally feasible to optimize the 0/1 loss directly.

Problem of gradient descent and gradient ascent

We need to choose the learning rate α rather carefully:

- Too small $\Rightarrow$ Slow convergence.

- Too large $\Rightarrow$ Overshoot and oscillation.



Newton–Raphson algorithm

- The Newton-Raphson algorithm: approximate the local shape of the loss function L as a quadratic function:

$$\vec{w}_{\text{new}} \leftarrow \vec{w} - H^{-1} \frac{\partial}{\partial \vec{w}} L(\vec{w}),$$

where H is the Hessian matrix of second derivatives:

$$H = \begin{pmatrix} \frac{\partial^2 L}{\partial w_0^2} & \frac{\partial^2 L}{\partial w_0 \partial w_1} & \dots & \frac{\partial^2 L}{\partial w_0 \partial w_d} \\ \frac{\partial^2 L}{\partial w_0 \partial w_1} & \frac{\partial^2 L}{\partial w_1^2} & \dots & \frac{\partial^2 L}{\partial w_1 \partial w_d} \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 L}{\partial w_d \partial w_0} & \frac{\partial^2 L}{\partial w_d \partial w_1} & \dots & \frac{\partial^2 L}{\partial w_d^2} \end{pmatrix}$$

- This is a **second-order** method, while gradient descent/ascent are **first-order** methods.

Generalized additive models

- As with regression, we can extend the MLE framework for logistic regression to arbitrary features (basis functions):

$$p(y = 1|\vec{x}) = \sigma(w_0 + \phi_1(\vec{x}) + \cdots + \phi_m(\vec{x})).$$

- **Example:** Quadratic logistic regression in 2D

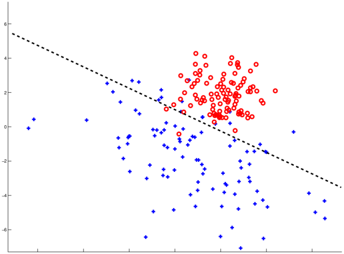
$$p(y = 1|\vec{x}) = \sigma(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2),$$

with quadratic decision boundary

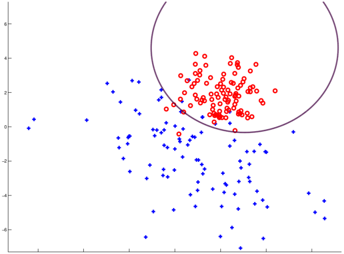
$$w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 = 0.$$

Generalized additive models

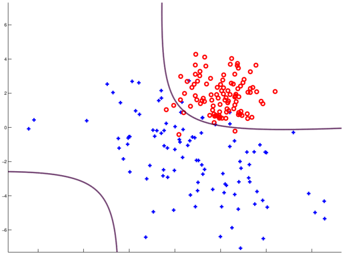
Linear



Quadratic



We can also include $x_1 x_2$:



Library for Logistic Regression

Examples

```
>>> from sklearn.datasets import load_iris
>>> from sklearn.linear_model import LogisticRegression
>>> X, y = load_iris(return_X_y=True)
>>> clf = LogisticRegression(random_state=0).fit(X, y)
>>> clf.predict(X[:2, :])
array([0, 0])
>>> clf.predict_proba(X[:2, :])
array([[9.8...e-01, 1.8...e-02, 1.4...e-08],
       [9.7...e-01, 2.8...e-02, ...e-08]])
>>> clf.score(X, y)
0.97...
```

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Next time

The course summary and practical questions for the final exam!