Logistic Regression

INSTRUCTOR: HONGJIE CHEN

MAY 31ST 2022

Problem Setting

- Learning $f : X \rightarrow Y$
- X is a real-valued vector $[X_1, X_2, \ldots, X_n]$
- Y is boolean

- Assume conditional independence given Y
- Model $P(X_i \mid Y = y_k)$ as Gaussian $\sim \mathcal{N}(\mu_{ik}, \sigma_i)$
- Model $P(Y)$ as Bernoulli $\sim \pi$

- What's the parametric form of $P(Y \mid X)$

Parametric form of $P(Y|X)$

$$P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$$

Law of total probability

Parametric form of $P(Y|X)$

$$P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1)+P(Y=0)P(X|Y=0)}$$

$$= \frac{1}{1+\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$

$$= \frac{1}{1+exp(\ln\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)})}$$

ln trick

$$= \frac{1}{1+exp(\ln\frac{1-\pi}{\pi}+\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)})}$$

Continue calculation 1

$$\sum_i ln\frac{P(X_i \mid Y=0)}{P(X_i \mid Y=1)} =$$

With $P(X_i = i \mid Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{(x-\mu_{ik})^2}{2\sigma_i^2}}$

Continue calculation 2

$$P(X_i = i \mid Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_i^2}}$$

$$\sum_i \ln \frac{P(X_i \mid Y=0)}{P(X_i \mid Y=1)}) = \sum_i \ln \frac{\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_{i0})^2}{2\sigma_i^2}}}{\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_{i1})^2}{2\sigma_i^2}}}$$

$$= \sum_i \ln e^{-(\frac{(x-\mu_{i0})^2}{2\sigma_i^2} - \frac{(x-\mu_{i1})^2}{2\sigma_i^2})}$$

$$= \sum_i - \frac{(x^2 - 2x\mu_{i0} + \mu_{i0}^2) - (x^2 - 2x\mu_{i1} + \mu_{i1}^2)}{2\sigma_i^2}$$

$$= \sum_i \frac{2(\mu_{i1} - \mu_{i0})x_i + \mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} = \sum_i \frac{\mu_{i1} - \mu_{i0}}{\sigma_i^2} x_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}$$

$w_i x_i$ Constant

Continue calculation 3

$$P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$$

$$= \frac{1}{1 + exp(ln\frac{1-\pi}{\pi} + \sum_i \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} + \sum_i \frac{\mu_{i1} - \mu_{i0}}{\sigma_i^2} x_i)}$$

$$= \frac{1}{1 + exp(w_0 + \sum_{i=1}^n w_i x_i)}$$

Optionally add $x_0 = 1$ to incoporate w_0 into the sum

Where
$$\begin{cases} w_0 & = \ln\frac{1-\pi}{\pi} + \sum_i \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} \\ w_i & = \frac{\mu_{i1} - \mu_{i0}}{\sigma_i^2} \end{cases}$$

Obtain Y directly from X with Parameters

$$P(Y=1|X) = \frac{1}{1+exp(w_0+\sum_{i=1}^{n} w_i x_i)}$$

$$\Rightarrow P(Y=0|X) = \frac{exp(w_0+\sum_{i=1}^{n} w_i x_i)}{1+exp(w_0+\sum_{i=1}^{n} w_i x_i)}$$

$$\Rightarrow \frac{P(Y=0|X)}{P(Y=1|X)} = exp(w_0+\sum_{i=1}^{n} w_i x_i)$$ **Compared with 1**

$$\Rightarrow ln\frac{P(Y=0|X)}{P(Y=1|X)} = w_0+\sum_{i=1}^{n} w_i x_i$$ **Compared with 0**

Predict $Y|X$ *in Short*

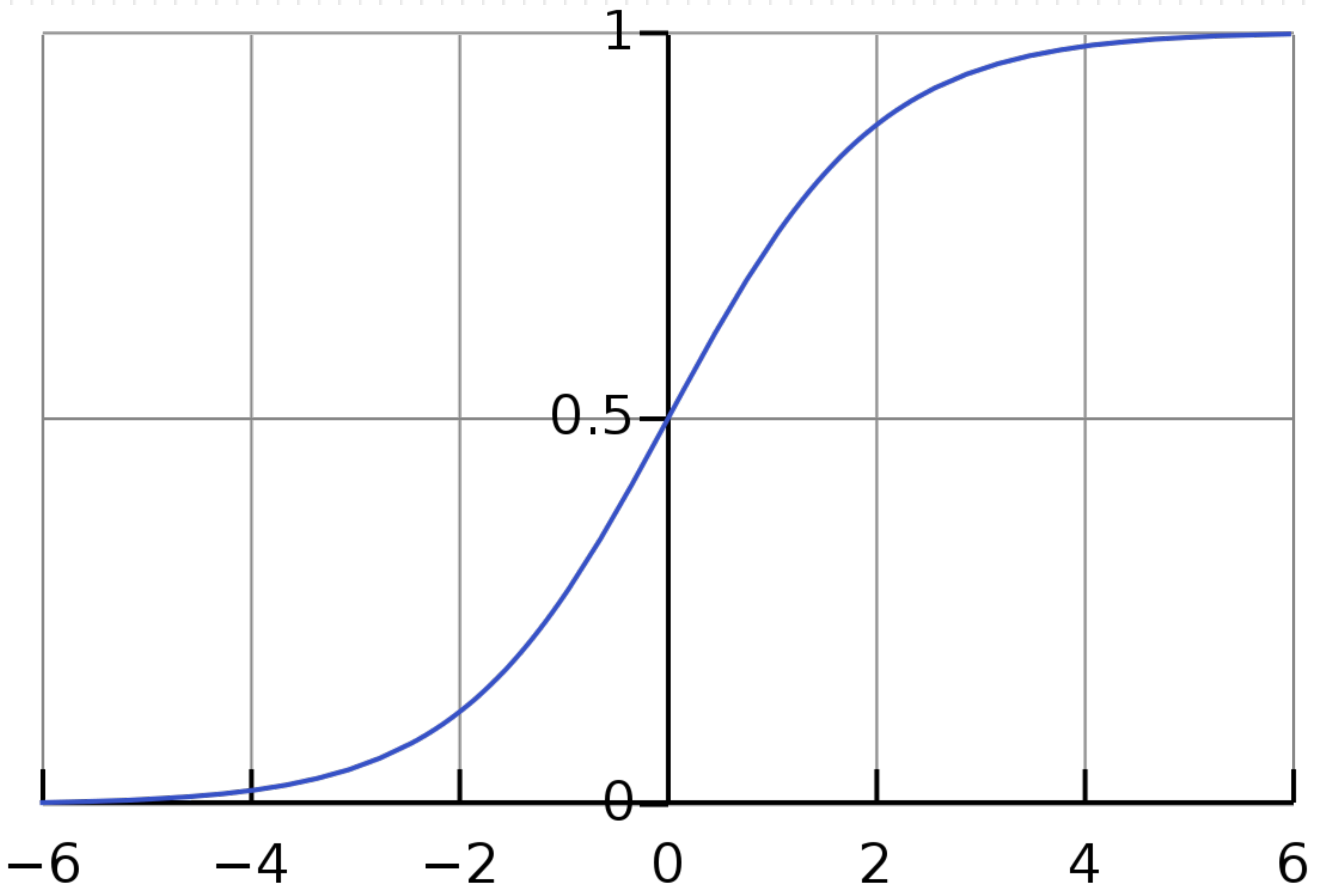
Calculate $w_0 + \sum_{i=1}^{n} w_i x_i$, predict $Y = 0$ if the result value is greater than 0, otherwise predict $Y = 1$

Logistic Regression (Generalized)

- Let's extend Y to contain more discrete values
  - Previously $Y \in \{0,1\}$, now $Y \in \{y_1, y_2, \ldots, y_R\}$
  - Learn $R - 1$ sets of weights
- For $k < R$: $P(Y = y_k | X) = \dfrac{exp(w_{k0} + \sum_{i=1}^{n} w_{ki} x_i)}{1 + \sum_{j=1}^{R-1} exp(w_{j0} + \sum_{i=1}^{n} w_{ji} x_i)}$
- For $k = R$: $P(Y = y_k | X) = \dfrac{1}{1 + \sum_{j=1}^{R-1} exp(w_{j0} + \sum_{i=1}^{n} w_{ji} x_i)}$

Logistic Function

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



Logistic Regression with MLE

- MLE?
- We have L training samples $\{(X^1, Y^1), \dots, (X^L, Y^L)\}$

$$W_{MLE} = \text{argmax}_W P((X^1, Y^1), \dots, (X^L, Y^L) \mid W)$$

$$= \text{argmax}_W \prod_l P((X^l, Y^l) \mid W)$$

- Have W to generate pairs of (X, Y)?

Logistic Regression MCLE

- Maximum Conditional Likelihood Estimation (MCLE)
- X is also conditioned

$$W_{MCLE} = \text{argmax}_W \prod_l P(Y^l | X^l, W)$$

Estimate MCLE

$$W_{MCLE} = \text{argmax}_W \prod_l P(Y^l | X^l, W)$$

- We are selecting good W (independent variable) to get highest $\prod_l P(Y^l | X^l, W)$ (dependent variable)

A function of W

- Again, assume Y is Boolean

$$f(W) = \ln \prod_l P(Y^l | X^l, W) = \sum_l \ln P(Y^l | X^l, W)$$

Express MCLE as a Function of W

$$f(W) = \sum_l \ln P(Y^l | X^l, W)$$

$$P(Y=0|X,W) = \frac{1}{1+exp(w_0 + \sum_{i=1}^n w_i x_i)}$$

With

$$P(Y=1|X,W) = \frac{exp(w_0 + \sum_{i=1}^n w_i x_i)}{1+exp(w_0 + \sum_{i=1}^n w_i x_i)}$$

$$f(W) = \sum_l Y^l \ln P(Y^l = 1 | X^l, W) + (1 - Y^l) \ln P(Y^l = 0 | X^l, W)$$

$$= \sum_l Y^l \ln \frac{P(Y^l = 1 | X^l, W)}{P(Y^l = 0 | X^l, W)} + \ln P(Y^l = 0 | X^l, W)$$

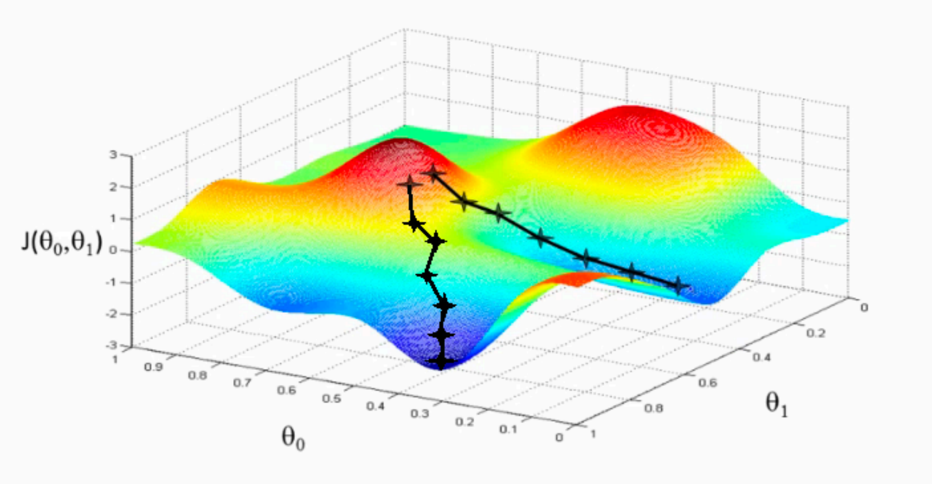
$$= \sum_l Y^l (w_0 + \sum_i^n w_i X_i^l) - \ln(1 + exp(w_0 + \sum_i^n w_i X_i^l))$$

Now we are taking the form and W is conditioned

VIRGINIA TECH

Gradient Ascent

- Gradient $\nabla f(\overrightarrow{w}) = [\frac{\partial f}{\partial w_0}, \frac{\partial f}{\partial w_1}, \ldots, \frac{\partial f}{\partial w_n}]$, is a vector
  - Parameter training rule: $\overrightarrow{w}^{(t+1)} \leftarrow \overrightarrow{w}^{(t)} + \eta \nabla f(\overrightarrow{w})$
  - View from one feature dimension $\Delta w_i = \eta \frac{\partial f}{\partial w_i}$



- Questions: What does η imply? What if we have a big η value.

VIRGINIA TECH

MCLE via Gradient Ascent

$$f(W) = \sum_l Y^l(w_0 + \sum_i^n w_i X_i^l) - \ln(1 + exp(w_0 + \sum_i^n w_i X_i^l))$$

$$\frac{\partial f(W)}{\partial w_i} = \sum_l X_i^l(Y^l - \hat{P}(Y^l = 1 \,|\, X^l, W))$$

- Gradient ascent algorithm: iterate until $\Delta w_i \ < \epsilon$
  - $\forall i: w_i \leftarrow w_i + \eta \sum_l X_i^l(Y^l - \hat{P}(Y^l = 1 \,|\, X^l, W))$
  - Incorporate w_0 with an assumed $X_0 = 1$
  - η is a hyperparameter: step size

Demo of Searching Best W

- https://yihui.org/animation/example/grad-desc/

- https://blog.skz.dev/gradient-descent

Batch v.s. Stochastic Gradient

- Batch gradient: use the entire training set D
  - Repeat until $\Delta w < \epsilon$
    - Compute the gradient: $\nabla f_D(\vec{w}) = [\frac{\partial f_D}{\partial w_0}, \frac{\partial f_D}{\partial w_1}, \ldots, \frac{\partial f_D}{\partial w_n}]$
    - Update parameters: $\vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} + \eta \nabla f_D(\vec{w})$
- Stochastic gradient: use a single sample $d \in D$ at a time
  - Repeat until $\Delta w < \epsilon$
    - Randomly Choose with replacement a training sample $d \in D$
    - Compute the gradient: $\nabla f_d(\vec{w}) = [\frac{\partial f_d}{\partial w_0}, \frac{\partial f_d}{\partial w_1}, \ldots, \frac{\partial f_d}{\partial w_n}]$
    - Update parameters: $\vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} + \eta \nabla f_d(\vec{w})$
- Which do we pick when $|D|$ is large?

Hyperparameters in Gradient-based Optimization

- Epoch:
  - An epoch referes to a full pass over the dataset
  - Each sample is used to update parameters once
  - The number of epochs is the number of full passes
  - Can work together with an early stopping strategy
- Batch size:
  - Batch size is the number of samples processed when the model is updated
  - An epoch can contain one or more batches

- For example, 10 training samples, 2 epochs, batch size as 4
  - 1st epoch
    - 1st iteration: a batch containg sample [1,2,3,4]
    - 2st iteration: a batch containg sample [5,6,7,8]
    - 3st iteration: a batch containg sample [9,10]
  - 2st epoch
    - 1st iteration: a batch containg sample [1,2,3,4]
    - 2st iteration: a batch containg sample [5,6,7,8]
    - 3st iteration: a batch containg sample [9,10]

Conduct experiments to decide hyperparameters

M(C)LE is good, what about MAP?

- Choose a prior

$$W_{MAP} = \text{argmax}_W P(W) \prod_l P(Y^l \mid X^l, W)$$

- Assume Gaussian prior: $W \sim \mathcal{N}(0, \sigma I)$

Weight Update with MAP

$$W_{MAP} = \text{argmax}_W P(W) \prod_l P(Y^l | X^l, W)$$

$$w_i \leftarrow w_i - \eta \lambda w_i + \eta \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1 | X^l, W))$$

Regularization term

- Avoids overfitting especially for sparse data
- Keeps weights near zero with prior

Naïve Bayes v.s. Logistic Regression

- Naïve Bayes
  - Assumption on $P(X|Y), P(Y)$
  - Estimates parameters of $P(X|Y), P(Y)$ from training data
  - Use Bayes rule to calculate $P(Y|X)$

- Logistic Regression
  - Assumption on $P(Y|X)$
  - Estimates parameters of $P(Y|X)$ directly from training data