Illustration: Theodore Modis

BBN406 Fundamentals of Machine Learning

Lecture 9: Logistic Regression Discriminative vs. Generative Classification



Erkut Erdem // Hacettepe University // Fall 2021

Last time... Naïve Bayes Classifier

Given:

- Class prior P(Y)
- d conditionally independent features X₁,...X_d given the class label Y
- For each X_i feature, we have the conditional likelihood $P(X_i|Y)$

Naïve Bayes Decision rule: $f_{NB}(\mathbf{x}) = \arg \max_{y} P(x_1, \dots, x_d \mid y) P(y)$ $= \arg \max_{y} \prod_{i=1}^{d} P(x_i \mid y) P(y)$

Last time... Naïve Bayes Algorithm for discrete features

 $f_{NB}(\mathbf{x}) = \arg \max_{y} \prod_{i=1}^{n} P(x_i|y)P(y)$ We need to estimate these probabilities!

Estimators
For Class Prior
$$\widehat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$
For Likelihood $\widehat{P}(x_i, y)$
 $\widehat{P}(y) = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$

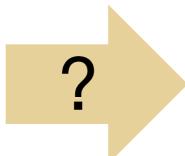
NB Prediction for test data:
$$X = (x_1, \dots, x_d)$$
 $Y = \arg \max_y \hat{P}(y) \prod_{i=1}^d \frac{\hat{P}(x_i, y)}{\hat{P}(y)}$

Last time... Text Classification

MEDLINE Article

RAINING PROVIDER	
	NY AND AND AND
	n aphasia: Plausibility judgments ubject sentences
	rger, ⁶ Duniel S. Junaféry, ⁶ Elizabeth Elder, ⁶ I L. Halland Audrey ⁴
 annova y manufactor and annova y manufactor and annova y manufactor annova y manufactor annova y manufactor a manufactor a	
United	
namenten ar vente mente fai latin fan de far af de main arts. In de namente Transfordings region (na Paramini fam) 2013 Bantin (na Alfright marvel)	en synthesely inder the neutron is study any in and inde 7 wines frequency and indea trians
	the factors giving rise to the greater difficulty of paratre
1. Introduction	strongend to carbon. Although the precise definition of
The simplety of "beneficial form," or "beneficial	managerie is an an a consequence of plant definition of manageries in the second seco
The simplify of "assessed form," or "assessed and units," for more and and pointed as as often have taken as at function in the assesses	manuality between jarray, lot is hence lines, 1997, constants who as proving other coult is to provide who share before adjust
The stopping of "beneficial lows," as "beneficial and solid," for neural and optimal competitions in order to a calculate the term patient competitions for and the same, as the term patient on the March (200), the order patient of constant on the March (200), the order patient of constant	manualities between latering, bein & horses from, 1600, secondaries with an provide with result to be transmitter with the participation reprint Conference approximates from the second active meta-instant with the second second test. Evider the
The simplify of "sectors lots," is "sectors and still," to sector out system surprising in the late two sides a sufficient is the sector surprising finance fiberation is being string and by Marc (200), it is string a local sector particu- tion have make application. (These statistics of "sectors" sector spin-size, it fibers statistics of "sectors" setting and the sector sectors.	- consistently is according to exp. Let in it. Response Hence, 1997, constraints with an providy white cond to be investigated with scheme to fairly depen- reported Dislargers arguments. Dashpits of comes- try provided to the scheme to be scheme to be scheme.
The simplify of "sectors later," or "sectors and still," for sector and spins surgetimizes a class have been as allocated in the sources surgetimes. Insector, its sector, a backet points on by Marc 2020, for which and the sources are been explored. These distances of version larget points and others and the back and reals approximate for these distances for agreent to be able to be distance and were to be to point in the distance of watering another to be to be able to be able to be distances.	maintaining is conserved interpret, family in the Response Harvey, 1983, summarize transition are provided in the internal on the interaction rates, where produce a depen- reporter the interpret regression. The interface of inter- action retains interactions. The interface of the inter- temporter and the interpret interaction. Under the mainterment is allowed provide an end of the inter- ient on the interactions. In they are allowed to mainter in the interpret in the product Dependence of the interface index of the product Dependence.
The straphicy of "horizontal large," as "horizontal and strapp" for several and application competition in a scalar large shown as an allocating to the several competition. Instance, it assumes a scalar particular to the Marcel Scalar and Scalar and Scalar barry Marcel Scalar and Scalar and Scalar barry Marcel Scalar and Scalar and Scalar barry and the application of second several barry and the attraction of second several to the start for the scalar and several to the scalar barry and the second several large scalar. Scalar and the second several large.	measure of presented interpy, form 4 Response Hanny, 1987, secondaries with any previse prob- tants in the transmission with any previse prob- senter the interpret regression. Reserve at a conse- cetion methods with the next and these. Duries in methods without any pro- metor without any pro- mains in the interpret of the pro- mains in the object of the promote the trans- tise pro- tom of the object of the promote the trans-
The simplify of "sectoral lots," is "sectoral and exists," for sector and sphere experiments in the later is the sector sector sector computation fraction. In the sector points in the later of the sector sector is the sector sector sector. These detections of the "sector discrete" point sector discrete sectors for sphere of the sector sector discrete sec- tor sphere is the sector of the sector discrete set is the instance of the sector discrete sec- tor sphere sector is the sector of the sector discrete like, here all, here all sector is the sector sector product the Digits is the sphere is the sector and the sector is the sector of the sector is the sector is the test of the sector of the sector.	consistency is the mean interpret for the A for present future, 1983, constraints on the original prob- result of the transition of the straint of the straint systems is the integrate approximation. Early in all const- nets on the straint of the straint of the straint of the constraints and the straints of the straints. The straints is straints and the straints of the straints of the constraints and the straints of the straints. The straints constraints and the straints of the straints of the straints of the straints of the straints of the straints of the straints and the straints of the straints of the straints and the straints of the straints of the straints and the straints of the straints of the straints in the straints of the straints of the straints of the straints in the straints of the straints of the straints.
	measure of presented interpy, form & Response Harrey, 1987, secondaries with any previse inde- tions in the transmission with any previse inde- tions in the transmission. Reserves all courses active exclusions and the latent has been based on membraneous analysis are and any based on the exclusion of the course of the State of the state of the exclusion of the system of the state of the state of the state of the system of the state of the state of the state of the system of the state of the state of the state of the system of the state of the state of the state of the system of the state of th

tion," for means that an analogue to difficulty should very with the balad bias of the



MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology

How to represent a text document?

Last time... Bag of words model

Typical additional assumption:

Position in document doesn't matter:

 $P(X_i = x_i | Y = y) = P(X_k = x_i | Y = y)$

- "Bag of words" model order of words on the page ignored
 The document is just a bag of words: i.i.d. words
- Sounds really silly, but often works very well!
- \Rightarrow K(5000-1) parameters to estimate

The probability of a document with words $x_1, x_2, ...$

$$\prod_{i=1}^{LengthDoc} P(x_i|y) = \prod_{w=1}^{W} P(w|y)^{count_w}$$

Last time... What if features are continuous?

e.g., character recognition: X_i is intensity at ith pixel



Gaussian Naïve Bayes (GNB): $P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$

Different mean and variance for each class k and each pixel i.

Sometimes assume variance

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

$$\widehat{\mu}_{MLE} = \frac{1}{N} \sum_{j=1}^{N} x_j$$

$$\hat{\sigma}_{unbiased}^2 = \frac{1}{N-1} \sum_{j=1}^N (x_j - \hat{\mu})^2$$

Logistic Regression

Recap: Naïve Bayes • NB Assumption: $P(X_1...X_d|Y) = \prod_{i=1}^d P(X_i|Y)$

- NB Classifier: $f_{NB}(x) = \arg \max_{y} \prod_{i=1}^{d} P(x_i|y)P(y)$
- Assume parametric form for $P(X_i|Y)$ and P(Y)
 - Estimate parameters using MLE/MAP and plug in

Gaussian Naïve Bayes (GNB)

- There are several distributions that can lead to a linear boundary.
- As an example, consider Gaussian Naïve Bayes:

 $Y \sim \text{Bernoulli}(\pi)$

$$P(X_i|Y=y) = \frac{1}{\sqrt{2\pi\sigma_{i,y}^2}} e^{\frac{-(X_i - \mu_{i,y})^2}{2\sigma_{i,y}^2}}$$

Gaussian class conditional densities

What if we assume variance is independent of class, i.e. $\sigma_{i,0}^2 = \sigma_{i,1}^2$

GNB with equal variance is a Linear Classifier!

$$P(X_i|Y=y) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(X_i - \mu_{i,y})^2}{2\sigma_i^2}}$$

Decision boundary:

$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=1)$$
$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(X_i|Y=1)P(X_i|Y=0)$$
$$\log \frac{P(Y=0)\prod_{i=1}^{d} P(X_i|Y=0)}{P(Y=1)\prod_{i=1}^{d} P(X_i|Y=1)} = \log \frac{1-\pi}{\pi} + \sum_{i=1}^{d} \log \frac{P(X_i|Y=0)}{P(X_i|Y=1)}$$

GNB with equal variance is a Linear Classifier!

$$P(X_i|Y=y) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(X_i - \mu_{i,y})^2}{2\sigma_i^2}}$$

Decision boundary:

$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=1)$$
$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=0)$$
$$\log \frac{P(Y=0)\prod_{i=1}^{d} P(X_i|Y=0)}{P(Y=1)\prod_{i=1}^{d} P(X_i|Y=1)} = \log \frac{1-\pi}{\pi} + \sum_{i=1}^{d} \log \frac{P(X_i|Y=0)}{P(X_i|Y=1)}$$

$$\frac{P(Y=0)\prod_{i=1}^{d}P(X_{i}|Y=0)}{P(X|Y=1)} = \frac{P(Y=0)\prod_{i=1}^{d}P(X_{i}|Y=0)}{P(X|Y=1)} = \log_{i}\frac{1-\pi}{\sigma_{i}^{2}} + \sum_{i=1}^{d}\log\frac{P(X_{i}|Y=0)}{\sigma_{i}^{2}} + \sum$$

GNB with equal variance is a Linear Classifier!

$$P(X_i|Y=y) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(X_i - \mu_{i,y})^2}{2\sigma_i^2}}$$

Decision boundary:

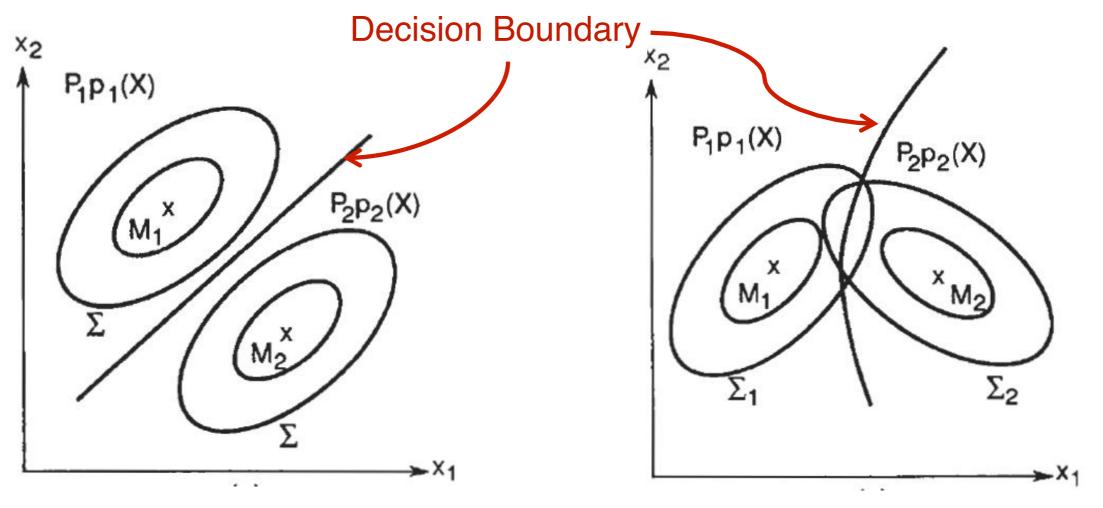
$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=1)$$
$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=0)$$
$$\log \frac{P(Y=0)\prod_{i=1}^{d} P(X_i|Y=0)}{P(Y=1)\prod_{i=1}^{d} P(X_i|Y=1)} = \log \frac{1-\pi}{\pi} + \sum_{i=1}^{d} \log \frac{P(X_i|Y=0)}{P(X_i|Y=1)}$$

$$\log \frac{P(Y=0) \prod_{i=1}^{d} P(X_i | Y=0)}{P(Y_{\log} \frac{1}{2}) \prod_{\pi} \frac{\pi}{i} \frac{d}{i} \sum_{i} \frac{P(X_i | \overline{Y} = 0)}{i} \sum_{i} \frac{1}{2\sigma_i^2} \frac{1}{2\sigma_i^2} \sum_{i} \frac{1}{\sigma_i^2} \frac{1}{\sigma_i^2} \sum_{i=1}^{d} \log \frac{P(X_i | Y=0)}{\sigma_i^2} \sum_{i=1}^{d} \log \frac{P(X_i | Y=$$

Constant term

First-order term

Gaussian Naive Bayes (GNB)



$$X = (x_1, x_2)$$

$$P_1 = P(Y = 0)$$

$$P_2 = P(Y = 1)$$

$$p_1(X) = p(X|Y = 0) \sim \mathcal{N}(M_1, \Sigma_1)$$

$$p_2(X) = p(X|Y = 1) \sim \mathcal{N}(M_2, \Sigma_2)$$

Generative vs. Discriminative Classifiers

- Generative classifiers (e.g. Naïve Bayes)
 - Assume some functional form for P(X,Y) (or P(X|Y) and P(Y))
 - Estimate parameters of P(X|Y), P(Y) directly from training data
- But arg max_Y P(X|Y) P(Y) = arg max_Y P(Y|X)
- Why not learn P(Y|X) directly? Or better yet, why not learn the decision boundary directly?
- Discriminative classifiers (e.g. Logistic Regression)
 - Assume some functional form for P(Y|X) or for the decision boundary
 - Estimate parameters of P(Y|X) directly from training data

Regression vs. Classification

Regression estimates a continuous value
Classification predicts a discrete category

MNIST: classify hand-written digits (10 classes)

ImageNet: classify nature objects (1000 classes)



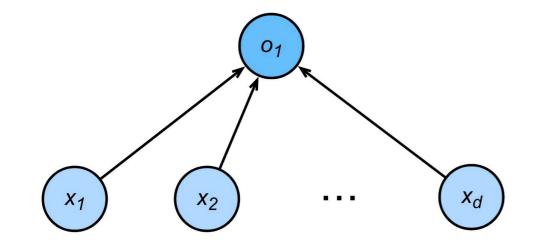


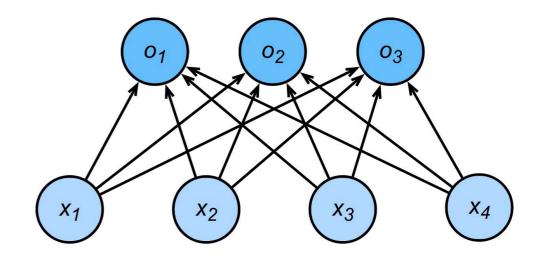
Regression

- Single continuous output
- Natural scale in
- Loss given e.g. in terms of difference y f(x)

Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence ...





Square Loss

• One hot encoding per class

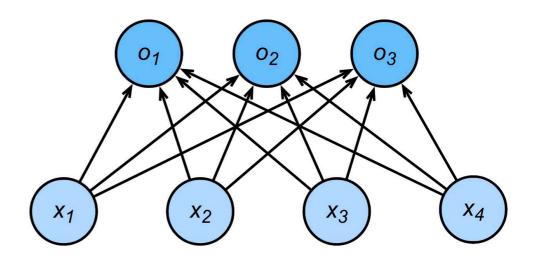
$$\mathbf{y} = [y_1, y_2, \dots, y_n]^{\mathsf{T}}$$
$$y_i = \begin{cases} 1 \text{ if } i = y \\ 0 \text{ otherwise} \end{cases}$$

- Train with squared loss
- Largest output wins

$$\hat{y} = \underset{i}{\operatorname{argmax}} o_i$$

Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence ...



Uncalibrated Scale

- One output per class
- Largest output wins

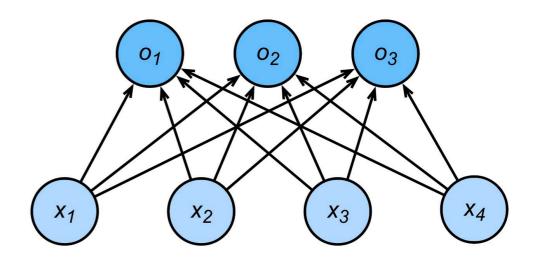
 $\hat{y} = \underset{i}{\operatorname{argmax}} o_i$

 Want that correct class is recognized confidently (large margin)

$$o_y - o_i \ge \Delta(y, i)$$

Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence ...



Calibrated Scale

• Output matches probabilities (nonnegative, sums to 1)

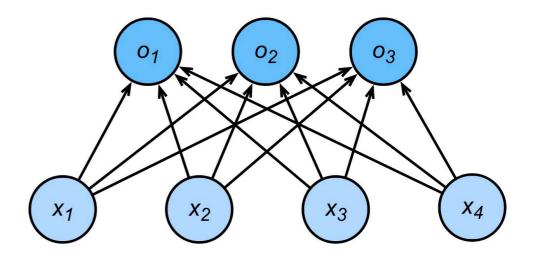
 $p(y \mid o) = \operatorname{softmax}(o)$ $= \frac{\exp(o_y)}{\sum_i \exp(o_i)}$

Classification

- Multiple classes, typically multiple outputs
- Score *should* reflect confidence ...

Negative log-likelihood

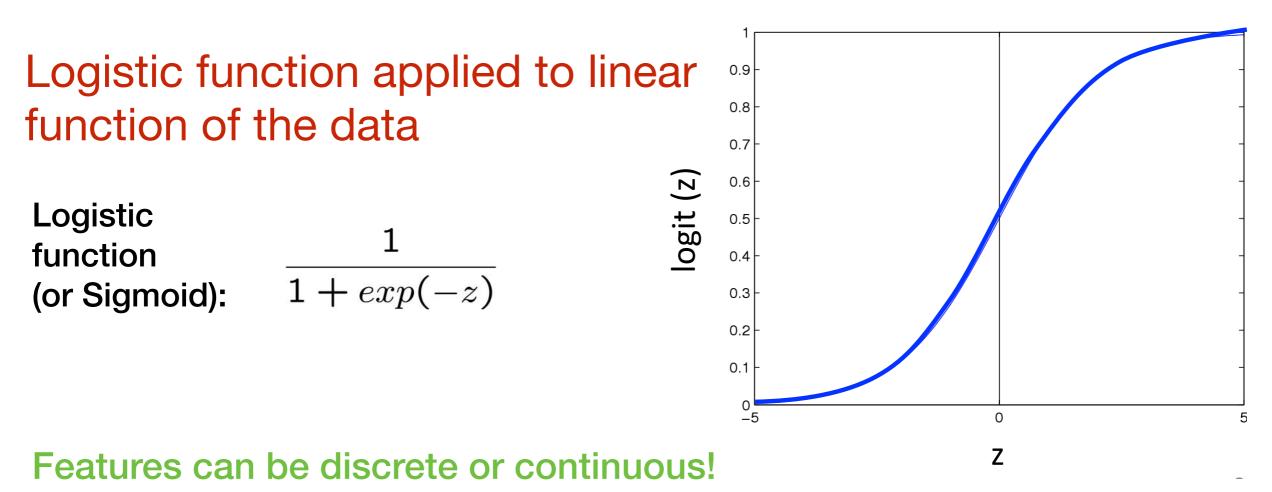
$$-\log p(y \mid o) = \log \sum_{i} \exp(o_{i}) - o_{y}$$



Logistic Regression

Assumes the following functional form for P(Y|X):

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$$



Logistic Regression is a Linear Classifier!

Assumes the following functional form for P(Y|X):

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

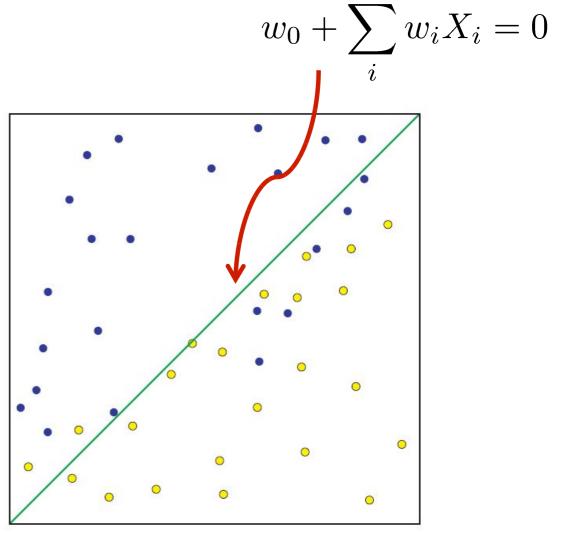
Decision boundary:

$$P(Y = 0|X) \stackrel{0}{\underset{1}{\gtrless}} P(Y = 1|X)$$

$$w_0 + \sum_i w_i X_i \stackrel{0}{\underset{1}{\gtrless}} 0$$

$$1$$

(Linear Decision Boundary)



Logistic Regression is a Linear Classifier!

Assumes the following functional form for P(Y|X):

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

$$\Rightarrow P(Y=0|X) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

$$\Rightarrow \frac{P(Y=0|X)}{P(Y=1|X)} = \exp(w_0 + \sum_i w_i X_i) \quad \stackrel{0}{\underset{1}{\gtrless}} \mathbf{1}$$
$$\Rightarrow \boxed{w_0 + \sum_i w_i X_i} \quad \stackrel{0}{\underset{1}{\gtrless}} \mathbf{0}$$

Logistic Regression for more than 2 classes

• Logistic regression in more general case, where $Y \in \{y_1, ..., y_K\}$

for kP(Y = y_k | X) = \frac{\exp(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i)}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^{d} w_{ji} X_i)}

for k=K (normalization, so no weights for this class)

$$P(Y = y_K | X) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^d w_{ji} X_i)}$$

Training Logistic Regression

We'll focus on binary classification:

$$P(Y = 0 | \mathbf{X}, \mathbf{w}) = \frac{1}{1 + exp(w_0 + \sum_i w_i X_i)}$$
$$P(Y = 1 | \mathbf{X}, \mathbf{w}) = \frac{exp(w_0 + \sum_i w_i X_i)}{1 + exp(w_0 + \sum_i w_i X_i)}$$

How to learn the parameters $w_0, w_1, ..., w_d$? Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$ $X^{(j)} = (X_1^{(j)}, ..., X_d^{(j)})$

Maximum Likelihood Estimates

$$\widehat{\mathbf{w}}_{MLE} = \arg \max_{\mathbf{w}} \prod_{j=1}^{n} P(X^{(j)}, Y^{(j)} | \mathbf{w})$$

Training Logistic Regression

We'll focus on binary classification:

$$P(Y = 0 | \mathbf{X}, \mathbf{w}) = \frac{1}{1 + exp(w_0 + \sum_i w_i X_i)}$$
$$P(Y = 1 | \mathbf{X}, \mathbf{w}) = \frac{exp(w_0 + \sum_i w_i X_i)}{1 + exp(w_0 + \sum_i w_i X_i)}$$

How to learn the parameters $w_0, w_1, ..., w_d$? Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$ $X^{(j)} = (X_1^{(j)}, ..., X_d^{(j)})$

Maximum Likelihood Estimates

$$\widehat{\mathbf{w}}_{MLE} = \arg \max_{\mathbf{w}} \prod_{j=1}^{n} P(X^{(j)}, Y^{(j)} | \mathbf{w})$$

But there is a problem... Don't have a model for P(X) or P(X|Y) — only for P(Y|X)

Training Logistic Regression

How to learn the parameters $w_0, w_1, ..., w_d$? Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$ $X^{(j)} = (X_1^{(j)}, ..., X_d^{(j)})$

Maximum (Conditional) Likelihood Estimates

$$\widehat{\mathbf{w}}_{MCLE} = \arg \max_{\mathbf{w}} \prod_{j=1}^{n} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})$$

Discriminative philosophy — Don't waste effort learning P(X), focus on P(Y|X) — that's all that matters for classification!

Expressing Conditional log Likelihood

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

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

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

Y can take only values 0 or 1, so only one of the two terms in the expression will be non-zero for any given *Y*¹

Expressing Conditional log Likelihood

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

Expressing Conditional log Likelihood P(Y=0|X) = P(Y=1|X) = P(Y=

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

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

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

Expressing Conditional log Likelihood P(Y=0|X) = P(Y=1|X) = P(Y=

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

$$\begin{aligned} \Psi(W) &= \sum_{l} Y^{l} \ln P(Y^{l} = 1 | X^{l}, W) + (1 - Y^{l}) \ln P(Y^{l} = 0 | X^{l}, W) \\ &= \sum_{l} Y^{l} \ln \frac{P(Y^{l} = 1 | X^{l}, W)}{P(Y^{l} = 0 | X^{l}, W)} + \ln P(Y^{l} = 0 | X^{l}, W) \\ &= \sum_{l} Y^{l} (w_{0} + \sum_{i}^{n} w_{i} X_{i}^{l}) - \ln(1 + \exp(w_{0} + \sum_{i}^{n} w_{i} X_{i}^{l})) \end{aligned}$$

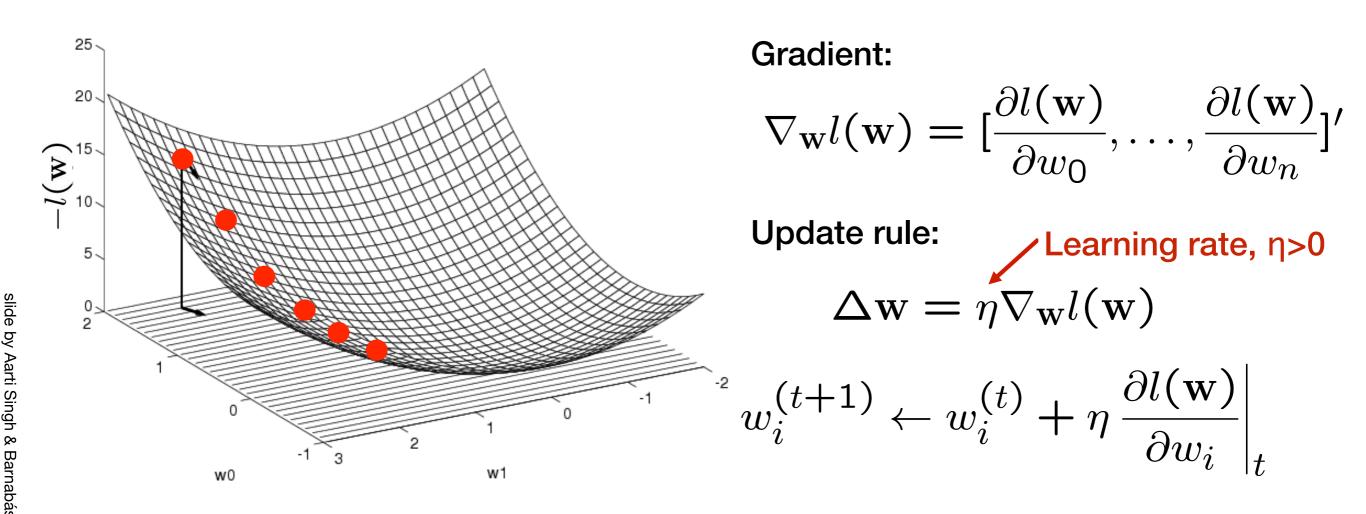
Maximizing Conditional log Likelihood

$$\max_{\mathbf{w}} l(\mathbf{w}) \equiv \ln \prod_{j} P(y^{j} | \mathbf{x}^{j}, \mathbf{w})$$
$$= \sum_{j} y^{j}(w_{0} + \sum_{i}^{d} w_{i} x_{i}^{j}) - \ln(1 + exp(w_{0} + \sum_{i}^{d} w_{i} x_{i}^{j}))$$

Bad news: no closed-form solution to maximize l(w)Good news: l(w) is concave function of w! concave functions easy to optimize (unique maximum)

Optimizing concave/convex functions

- Conditional likelihood for Logistic Regression is concave
- Maximum of a concave function = minimum of a convex function
 Gradient Ascent (concave)/ Gradient Descent (convex)



Gradient Ascent for Logistic Regression

Gradient ascent algorithm: iterate until change < ε

$$w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \sum_j [y^j - \hat{P}(Y^j = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})]$$

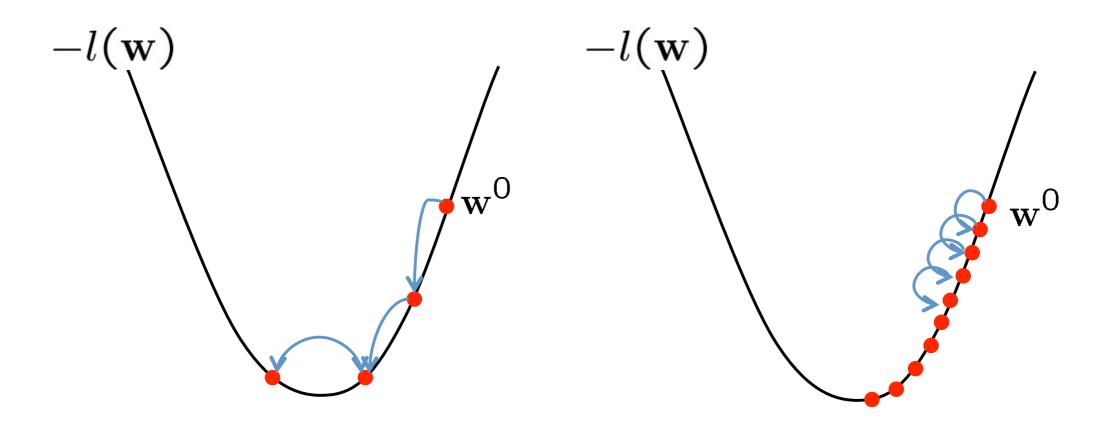
$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \sum_j x_i^j [y^j - \hat{P}(Y^j = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})]$$

repeat

Predict what current weight thinks label Y should be

- Gradient ascent is simplest of optimization approaches
 - e.g. Newton method, Conjugate gradient ascent, IRLS (see Bishop 4.3.3)

Effect of step-size η



Large $\eta \rightarrow$ Fast convergence but larger residual error Also possible oscillations

Small $\eta \rightarrow$ Slow convergence but small residual error

Set of Gaussian Naïve Bayes parameters (feature variance independent of class label)



Set of Logistic Regression parameters

- Representation equivalence
 - But only in a special case!!! (GNB with class-independent variances)
- But what's the difference???

Set of Gaussian Naïve Bayes parameters (feature variance independent of class label)



Set of Logistic Regression parameters

- Representation equivalence
 - But only in a special case!!! (GNB with class-independent variances)
- But what's the difference???
- LR makes no assumption about P(X|Y) in learning!!!
- Loss function!!!
 - Optimize different functions! Obtain different solutions

Consider Y Boolean, X_i continuous X=<X₁ ... X_d>

Number of parameters:

- NB: 4d+1 π , ($\mu_{1,y}$, $\mu_{2,y}$, ..., $\mu_{d,y}$), ($\sigma^2_{1,y}$, $\sigma^2_{2,y}$, ..., $\sigma^2_{d,y}$) y=0,1
- LR: d+1 w₀, w₁, ..., w_d

Estimation method:

- NB parameter estimates are uncoupled
- LR parameter estimates are coupled

Generative vs. Discriminative

[Ng & Jordan, NIPS 2001]

Given infinite data (asymptotically),

If conditional independence assumption holds, Discriminative and generative NB perform similar.

 $\epsilon_{\mathrm{Dis},\infty}\sim\epsilon_{\mathrm{Gen},\infty}$

If conditional independence assumption does NOT holds, Discriminative outperforms generative NB.

$$\epsilon_{\mathrm{Dis},\infty} < \epsilon_{\mathrm{Gen},\infty}$$

Generative vs. Discriminative

[Ng & Jordan, NIPS 2001]

Given finite data (n data points, d features),

$$\epsilon_{\mathrm{Dis},n} \leq \epsilon_{\mathrm{Dis},\infty} + O\left(\sqrt{\frac{d}{n}}\right)$$

 $\epsilon_{\mathrm{Gen},n} \leq \epsilon_{\mathrm{Gen},\infty} + O\left(\sqrt{\frac{\log d}{n}}\right)$

Naïve Bayes (generative) requires $n = O(\log d)$ to converge to its asymptotic error, whereas Logistic regression (discriminative) requires n = O(d).

Why? "Independent class conditional densities"

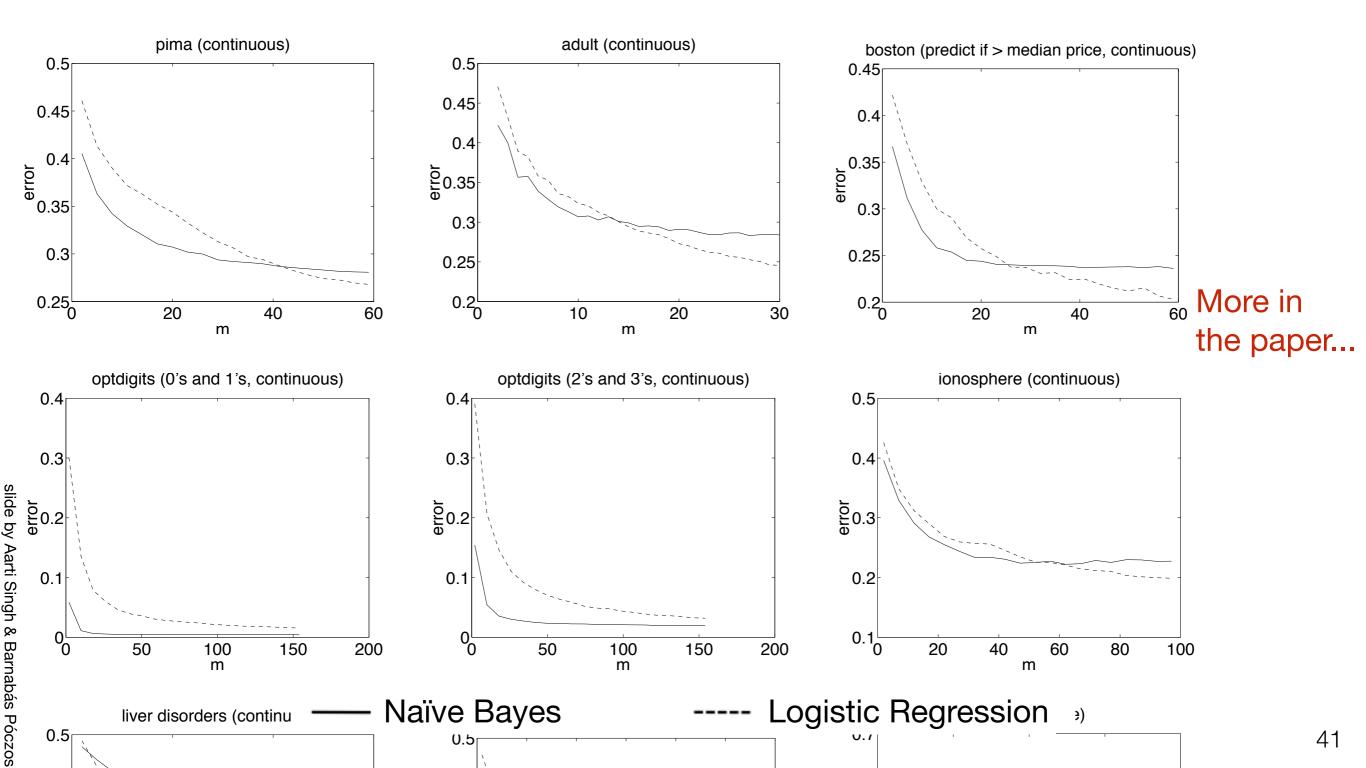
 parameter estimates not coupled – each parameter is learnt independently, not jointly, from training data.

Verdict

Both learn a linear decision boundary. Naïve Bayes makes more restrictive assumptions and has higher asymptotic error, BUT converges faster to its less accurate asymptotic error.

Experimental Comparison (Ng-Jordan'01)

UCI Machine Learning Repository 15 datasets, 8 continuous features, 7 discrete features



What you should know

- \cdot LR is a linear classifier
 - decision rule is a hyperplane
- LR optimized by maximizing conditional likelihood
 - no closed-form solution
 - concave ! global optimum with gradient ascent
- Gaussian Naïve Bayes with class-independent variances representationally equivalent to LR
 - Solution differs because of objective (loss) function
- In general, NB and LR make different assumptions
 - NB: Features independent given class! assumption on P(X|Y)
 - LR: Functional form of P(Y|X), no assumption on P(X|Y)
- Convergence rates
 - GNB (usually) needs less data
 - LR (usually) gets to better solutions in the limit

Next Lecture: Linear Discriminant Functions Perceptron