Illustration: Theodore Modis

BBN406 Fundamentals of Machine Learning

Lecture 9: Logistic Regression Discriminative vs. Generative Classification



Aykut Erdem // Hacettepe University // Fall 2019

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

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

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

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

····· d	B	ognition
		and the second second
Syntactic frame and verb bias in of undergoer-st	raphasia: Plausibility jud ubject sentences	gments
Seaanse Gulti,* Liee Mans,* Call Romaher Molly Reseas,* and	ger, ^b Daniel S. Jorafiky, ^b Elizab L. Hatland Audrep ^a	eth Elder,"
Advancements Advan	fantrige an, add ar, Austri fin, and ar, Panar AR, addr 1 May ann	
The analysis was presented to a track that a provide the second state of the second state of a second distribution of the second state of the second distribution of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of t	 A first "Associate" for the second sec	animo figurante las activa presente prese d'acchante presente presente
The only homogeneous latent for loss the equal process strength of a set flagsing of any 20 km strength in the strength of the set strength of the set strength of the strength of the set strength of the set strength of a set of the set of the set of the set of the set strength of the set strength of the set of the set strength of the set of the Set (the set of the set of the set of the set of the set of the set of the set of the set of the Set (the set of the set of the set of the Set (the set of the	• Address "According," https://www.according.com/ www.according.com/ www.according.com/ www.according.com/ by https://www.according.com/ www.according.com/ by https://www.according.com/ wwww.according.com/ www.according	enter Ayraco la otto poer poer al episora querc'a talan episor in a pint o poerc'a statu arcon esti actua
The only homogeneous lasts has been been equally record patients at a set flapping of any The bar water in a strength of the set of	• Allow "According, here," In any part or party of the state of the sector of the state of th	enter fyrneli rog of channel prog of channel providy when a providy me- represent a form
The only have part that have been been equal to serve attends the set flagsing of any 2000 the tensor photon and photon in the set of the set o	• All in "According, Intel" in an energy of a second se	enter fytuett prop d splane grand y take grand y take grand y take grand y take grand y take t grand y take t grand y take t grand y take t grand y take
The endy hanging in the latent into here have a par- ticities a second second second second second second second distribution of the second second second second second distribution of the second sec	A site "second limit" in an an except in the second second second process in the second second second process in the second s	entre fyraett produktioner produktioner produktioner produktioner produktioner ander af souther ander af souther and souther and souther produktioner and souther and souther
The only background is the first back for the second filter is a second system in the back for the second filter is a subgroup on the set of particle sectors. Using a se- tematic second second second second second second second second second second second second second second second second second secon	In the free state, in the free state as experimental to the process of the state	entre fyraets produktion a part produktion a part produktion a part produktion a part arter part a part part a
The story is a significant factor for long the long of the story of th	In the "Assessed from" to an our comparison of the second seco	enter former se alle present prese d'appart prese la grand prese la grand se prese l
The story is a single in the last in the last is a space of the story	In the Second Intel [®] is a new coupling to the second	enter fyrein i en eine afgeste present generally terles generally state generally state generally state and state afferter afferter in de terlester afferter in de terlester afferter a
The endpt transport is the first that the start the start fitter is every a start of particular that the start fitter is every a start of particular that the start fitter is every a start of particular that the start is a start of the start of the start of the start is a start of the start of the start of the start is a start of the start particular the start of the start of the start of the start particular the start of the start of the start of the start particular the start of the start of the start of the start particular the start of the start of the start of the start particular the start of the start of the start of the start particular the start of the start of the start of the start of the start of the start of the start of the start of the start is the start of the start of the start of the start of the start is the start of the start of the start of the start of the start of the start of the start of the start of the start of the start of the start of the start of the start of the	In the formula low " to array a couple in the result is a structure expering reason by the structure is a structure expering reason by the structure is the structure result is a property of the structure result is a structure by the structure is a structure result in the structure is a structure in the structure is a structure in the structure is a structure in the structure property for a structure is a structure is a structure in the structure is a structure in the structure is a structure in a structure is a structure in the structure is a structure in the structure is a structure in a structure is a structure in the structure is a structure is a structure in the structure is a structure is a structure in the structure is a structure in the structure is a structure in the a structure is structure in the structure is a structure in a structure is a structure in the structure is a structure in a structure is a structure in the structure is a structure in a structure is a structure in the structure in the structure is a structure is a structure in the structure in the structure is a structure in the structure is a structure in the structure is a structure in the structure in the structure in the structure is structure in the s	and the figure of the second s
The original program is then be been been equal to the energy of the original structure is the second distribution program is the distribution. They are been to active involves in the second structure is the second structure energy is the second structure is the second struc- ture active events in the second structure is the second struc- ture active events in the second structure is the second struc- ture active events in the second structure is the second struc- ture active events in the second structure is the second struc- ture active events in the second structure is the second struc- ture active events in the second structure is the second active events in the second structure is the second struc- ture active is the second structure is the second struc- ture active is the second structure is the second struc- ture active is the second structure is the structure is and a structure (Salaran et al. Salaran et al. Salaran is the second structure (Salaran et al. Salaran et al. Salaran is the second structure is the structure for structure is the second structure is the structure of the second structure. Earlies are activated as the second structure is the second activates of the second law of the second structure is the second structure (Salaran et al. Salaran et al. Salaran structure is the second structure is the second activates of the second structure is the second structure is the second structure activate structure is the second structure in the second structure structure (Salaran activates is the second provide structure structure (Salaran et al. Salaran activates is the second structure structure (Salaran et al. Salaran Salaran Salaran activates in the second structure structure (Salaran activates) for the second structure provide structure structure (Salaran et al. Salaran Salaran activates) is the second structure structure (Salaran et al. Salaran Salaran activates) is the second structure structure structure (Salaran activates) is the second structure of the second structure of the second structure	In the formula have been experimented as a second secon	and the figure of the second s
The order to explorate the term for the form that the explore transmission of the order to explore the part of the fiber term optimum is an effective term for the part of the term optimum is the order term of the the explore term of the term optimum is the term of the term of the part of terms of the term of the term of the term of the term of the term of the term of the term of the term of the term of the term of the term of the term of term of the term of the term of the term of term of terms of term of the term of the term of term of term of terms of term of term of term of the term of term of terms of terms of term of term of term of terms of terms of terms of terms of term of terms of terms of terms of terms of terms of terms of term of terms o	In the formula low the mean exception of the second low of the second	and the figure of the second s



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)

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

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

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 $c_{i,0}^2$ as $\sigma_{i,1}^2$ i.e. $\sigma_{i,0}^2 = \sigma_{i,1}^2$

 $_{i,y})^2$

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:

$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=1)$$
$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=0)$$

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

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

 $P(Y=0)P(X|Y=0) = 1 - \pi + \sum_{i=1}^{N} \mu_{i,1}^2 - \mu_{i,0}^2 + \sum_{i=1}^{N} \mu_{i,0}^2 - \mu_{i,1} + \sum_{i=1}^{N} \mu_{i,0}^2 - \mu_{i,0}^2 + \sum_{i=1}^{N} \mu_{i,0}^2$

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:

$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=1)$$
$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=0)$$

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

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

 $P(Y=0)P(X|Y=0) = 1 - \pi + \sum_{i=1}^{N} \mu_{i,1}^2 - \mu_{i,0}^2 + \sum_{i=1}^{N} \mu_{i,0}^2 - \mu_{i,1} + \sum_{i=1}^{N} \mu_{i,0}^2 - \mu_{i,0}^2 + \sum_{i=1}^{N} \mu_{i,0}^2$

GNB with equal variance is a Linear Classifier! $P(X_i|Y=y)$ $P(X_i|Y=y) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(X_i - \mu_{i,y})^2}{2\sigma_i^2}}$

Decision boundary:

$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=1)$$
$$\prod_{i=1}^{d} P(X_i|Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=0)P(Y=0) = \prod_{i=1}^{d} P(X_i|Y=1)P(Y=0)P(Y=$$

 $\log \frac{P(Y=0) \prod_{i=1}^{a} P(X_i | Y=0)}{P(Y=1) \prod_{i=1}^{d} P(X_i | Y=1)} = \log \frac{1-\pi}{\pi} + \sum_{i=1}^{a} \log \frac{P(X_i | Y=0)}{P(X_i | Y=1)}$ $\log \frac{P(Y=0) \prod_{i=1}^{d} P(X_i | Y=0)}{P(Y_{\log} \frac{1}{4}) \prod_{i=1}^{\pi} \int_{i=1}^{d} \sum_{i} \left(\frac{\mu_{i,1}^2}{X_i} | \overline{Y} | \frac{\overline{Y} \mu_{i,0}^2}{2\sigma_i^2} 1 \right)}{2\sigma_i^2} \sum_{i=1}^{d} \frac{\log_{i} \frac{1-\pi}{\pi}}{\sigma_i^2} X_i \sum_{i=1}^{d} \log \frac{P(X_i | Y=1)}{2\sigma_i^2} \sum_{i=1}^{d} V_i | \overline{Y} | \frac{\overline{Y} \mu_{i,0}}{2\sigma_i^2} \sum_{i=1}^{d} V_i | \frac{\overline{Y}$ $\int_{2}^{\infty} P(Y=0)P(X|Y \text{ Constant term } -\pi + \text{First-order term } + \sum \mu_{i,0} - \mu_{i,1} + \sum \mu_{i,0} - \mu_{i,0} + \sum \mu_{i,0} - \mu_{i,1} + \sum \mu_{i,0} + \sum \mu_{i,0} - \mu_{i,0} + \sum \mu_{i,0} + \sum$

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

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:

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

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

(Linear Decision Boundary)

i



Logistic Regression is a Linear Classifier!

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

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

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

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

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) =

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

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

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

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

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

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

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 $< \epsilon$

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

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

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.

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

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

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

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