Learning as Inference

DD2421

Bob L. T. Sturm

Check your understanding!

Before I took a COVID test, the doctor said 99% of the people in the area have COVID, and 90% of those with COVID are testing positive. A few days later the doctor called and said my test was positive, and that the probability I have COVID given this positive test is p% — I can't remember because I was in shock. Find the minimum value of p such that I can compute the probability I got a positive test but don't have COVID, and then compute the maximum probability I don't have COVID given my positive test.

Check your understanding!

Before I took a COVID test, the doctor said 99% of the people in the area have COVID, and 90% of those with COVID are testing positive. A few days later the doctor called and said my test was positive, and that the probability I have COVID given this positive test is p% — I can't remember because I was in shock. Find the minimum value of p such that I can compute the probability I got a positive test but don't have COVID, and then compute the maximum probability I don't have COVID given my positive test.

Define:

1. $+$ is "positive test"
2. C is "have COVID"
3. $\neg C$ is "does not have COVID"

Check your understanding!

Before I took a COVID test, the doctor said 99% of the people in the area have COVID, and 90% of those with COVID are testing positive.

Translating:

- $P[C] = 0.99$ (prior)
- $P[+|C] = 0.9$ (likelihood)

Check your understanding!

The doctor called and said my test was positive, and that the probability I have COVID given this positive test is p% — I can't remember because I was in shock.

Translating:

- OMG I have "$+$" (evidence)
- $P[C|+] = p$ (posterior)

Check your understanding!

Find the minimum value of $p \equiv P[C|+]$ such that I can compute the probability I got a positive test but don't have COVID

Translating:

- I want to find $P[+|\neg C]$.
- I also know $P[C|+]$ is a probability, and so its value must be in a range restricted by the axioms of probability.

Check your understanding!

Find the minimum value of $p \equiv P[C|+]$ such that I can compute the probability I got a positive test but don't have COVID
Translating:

- I want to find $P[+|\neg C]$.
- I also know $P[C|+]$ is a probability, and so its value must be in a range restricted by the axioms of probability.

Using Bayes':

$$P[+|\neg C] = \frac{P[\neg C|+]P[+]}{P[\neg C]} = \frac{(1 - P[C|+])P[+]}{1 - P[C]} = \frac{(1-p)P[+]}{1 - P[C]}$$

since $P[\neg C|+] = 1 - P[C|+]$, and $P[\neg C] = 1 - P[C]$ as there are only two possibilities. We need to find $P[+]$.

Check your understanding!

We know

$$P[+|C] = \frac{P[C|+]P[+]}{P[C]}$$

and so solving for $P[+]$

$$P[+] = \frac{P[+|C]P[C]}{P[C|+]} = \frac{P[+|C]P[C]}{p}$$

Check your understanding!

We know:

$$P[+|\neg C] = \frac{(1-p)P[+]}{1-P[C]}$$

and we have just found

$$P[+] = \frac{P[+|C]P[C]}{p}$$

Substituting the latter into the former produces

$$\begin{aligned} P[+|\neg C] &= \frac{(1-P[C|+])}{1-P[C]}\frac{P[+|C]P[C]}{P[C|+]} \\ &= \frac{P[C]}{1-P[C]}\frac{(1-P[C|+])}{P[C|+]}P[+|C] \\ &= \frac{P[C]}{1-P[C]}\frac{(1-p)}{p}P[+|C]. \end{aligned}$$

Check your understanding!

Our crowning achievement:

$$P[+|\neg C] = \frac{P[C]}{1-P[C]}\frac{(1-p)}{p}P[+|C].$$

The left hand side must obey the axioms of probability, which means $0 \le P[+|\neg C] \le 1$. So

$$0 \le \frac{P[C]}{1-P[C]}\frac{(1-p)}{p}P[+|C] \le 1$$
$$0 \le \frac{(1-p)}{p} \le \frac{1-P[C]}{P[C]}\frac{1}{P[+|C]}$$
$$0 \le \frac{(1-p)}{p} \le \frac{1/100}{99/100}\frac{1}{9/10} \to 891/901 \le p \le 1$$

Check your understanding!

The minimum value of p such that I can compute the probability I got a positive test but don't have COVID:

$$p \geq 891/901$$

The maximum probability I don't have COVID given my positive test is thus:

$$P[\neg C|+] = 1 - P[C|+] = 1 - p \leq 1 - 891/901 = 10/901.$$

Outline

Outline

Probabilistic Classification and Regression

- In both cases we compute the posterior

$$Pr(y \,|\, X = x) = \frac{Pr(x \,|\, Y = y) Pr(Y = y)}{Pr(X = x)}$$

Probabilistic Classification and Regression

- In both cases we compute the posterior

$$Pr(y \,|\, X = x) = \frac{Pr(x \,|\, Y = y) Pr(Y = y)}{Pr(X = x)}$$

- Classification: Y is discrete, finite
- Regression: Y is continuous

Until now we assumed we knew:

- $Pr(Y = y) \equiv Pr(y) \leftarrow$ *Prior*
- $Pr(x \,|\, Y = y) \equiv Pr(x|y) \leftarrow$ *Likelihood*
- $Pr(X = x) \equiv Pr(x) \leftarrow$ *Evidence*

Probabilistic Classification and Regression

- In both cases we compute the posterior

$$Pr(y \,|\, X = x) = \frac{Pr(x \,|\, Y = y) Pr(Y = y)}{Pr(X = x)}$$

- Classification: Y is discrete, finite
- Regression: Y is continuous

Until now we assumed we knew:

- $Pr(Y = y) \equiv Pr(y) \leftarrow$ *Prior*
- $Pr(x \,|\, Y = y) \equiv Pr(x|y) \leftarrow$ *Likelihood*
- $Pr(X = x) \equiv Pr(x) \leftarrow$ *Evidence*

How can we obtain these distributions from data?

Learning as Inference

Given:

- the training data $\mathcal{D} = \{(\mathbf{x}, y)_1, (\mathbf{x}, y)_2, \ldots, (\mathbf{x}, y)_N\}$
- a new observation $\mathbf{x}$

Estimate the posterior probability of y:

$$Pr(y|\mathbf{x}, \mathcal{D})$$

Discriminative vs Generative Models

Discriminative modeling:

- This models $Pr(y|\mathbf{x}, \mathcal{D})$ directly
- examples: logistic regression

Generative modeling:

- This models $Pr(\mathbf{x}|y, \mathcal{D})$
- example: Naive Bayes

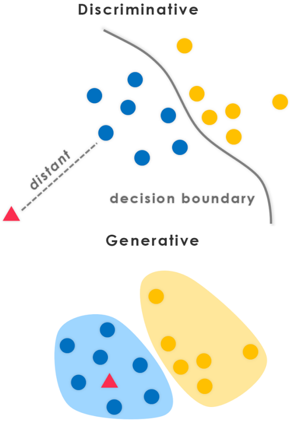


Figure from Nguyen *et al.* 2015.

Parametric vs Non-parametric Inference

$$Pr(y|\mathbf{x}) = Pr(y|\mathbf{x}, \theta)$$

The distribution is characterized by parameters θ.

Parametric vs Non-parametric Inference

$$Pr(y|\mathbf{x}) = Pr(y|\mathbf{x}, \theta)$$

The distribution is characterized by parameters θ.

Parametric Inference:

- Estimate θ using $\mathcal{D}$
- Compute $Pr(y|\mathbf{x}, \hat{\theta})$ to make inference.

Learning corresponds to estimating θ

Parametric vs Non-parametric Inference

$$Pr(y|\mathbf{x}) = Pr(y|\mathbf{x}, \theta)$$

The distribution is characterized by parameters θ.

Parametric Inference:

- Estimate θ using $\mathcal{D}$
- Compute $Pr(y|\mathbf{x}, \hat{\theta})$ to make inference.

Learning corresponds to estimating θ

Non-Parametric Inference:

- Estimate $Pr(\theta|\mathcal{D})$
- Compute $Pr(y|\mathbf{x}, \mathcal{D})$ from $Pr(y|\mathbf{x}, \theta, \mathcal{D})Pr(\theta|\mathcal{D})$ by marginalizing out θ

The number of parameters can grow with the data!

Three Approaches

Parametric inference:

- Maximum Likelihood (ML) Estimation (today)
- Maximum A Posteriori (MAP) Estimation (next time)

Non-parametric inference:

- Bayesian methods (a little today and the rest next time)

Fundamental Assumption: i.i.d.

Observations are independent and identically distributed (i.i.d.):

$$\mathcal{D} = \{\mathbf{o}_1, \ldots, \mathbf{o}_N\}, \mathbf{o}_i = (\mathbf{x}, y)_i$$

The likelihood of the whole data set can be factorized:

$$Pr(\mathcal{D}) = Pr(\mathbf{o}_1, \ldots, \mathbf{o}_N) = \prod_{i=1}^{N} Pr(\mathbf{o}_i)$$

Taking the log creates the *log-likelihood*:

$$\log Pr(\mathcal{D}) = \sum_{i=1}^{N} \log Pr(\mathbf{o}_i)$$

Outline

Maximum Likelihood (ML) Estimate

$$Pr(\mathbf{x}|y) \equiv Pr(\mathbf{x}|y,\theta) \quad \text{or} \quad Pr(y|\mathbf{x}) \equiv Pr(y|\mathbf{x},\theta)$$

Find the parameter values that make the data most likely.

- *ML optimality* is defined as maximizing the likelihood of $\mathcal{D}$:

$$\theta_{\text{ML}} = \arg\max_{\theta} P(\mathcal{D}|\theta) = \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

- We can then approximate distributions given the data:

$$Pr(\mathbf{x}|y,\mathcal{D}) \approx Pr(\mathbf{x}|y,\theta_{\text{ML}}) \quad \text{or} \quad Pr(y|\mathbf{x},\mathcal{D}) \approx Pr(y|\mathbf{x},\theta_{\text{ML}})$$

Probabilistic Linear Regression

Model (deterministic):

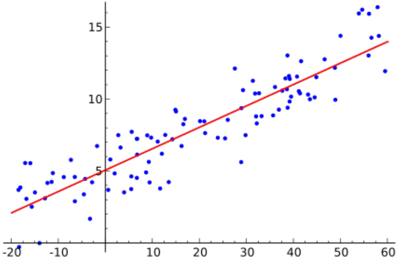
$$y = \mathbf{w}^T \mathbf{x} + \epsilon$$

But now:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Therefore:

$$Y|X \sim \mathcal{N}(\mu_Y(\mathbf{x}), \sigma_Y^2(\mathbf{x})) = \mathcal{N}(\mathbf{w}^T \mathbf{x}, \sigma^2)$$



Learning: find $\mathbf{w}$ that maximizes $Pr(y|\mathbf{x}, \mathbf{w}, \sigma^2)$

Maximize the posterior directly $\implies$ discriminative method

MLE for Probabilistic Linear Regression

$$\begin{aligned}
\log Pr(y|\mathbf{x}, \mathbf{w}, \sigma^2) &= \log \prod_i Pr(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) \\
&= \sum_i \log Pr(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) \\
&= \sum_i \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2}} \right] \\
&= \sum_i \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2} \right]
\end{aligned}$$

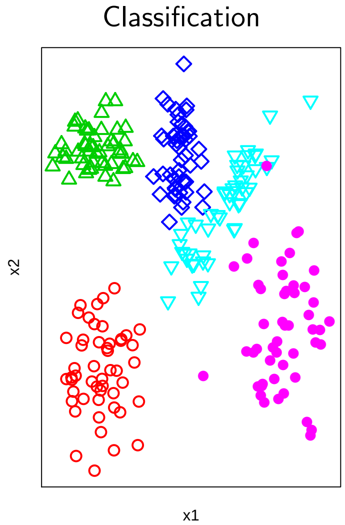
MLE for Probabilistic Linear Regression

$$\begin{aligned}
\log Pr(y|\mathbf{x}, \mathbf{w}, \sigma^2) &= \log \prod_i Pr(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) \\
&= \sum_i \log Pr(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) \\
&= \sum_i \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2}} \right] \\
&= \sum_i \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2} \right]
\end{aligned}$$

$$\arg\max_{\mathbf{w}} Pr(y|x, \mathbf{w}, \sigma^2) = \arg\min_{\mathbf{w}} \sum_i (y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

NEAT-O! Choosing parameters that maximize $Pr(y|x, \mathbf{w}, \sigma^2) \equiv$ minimizing mean square error! (in this case)

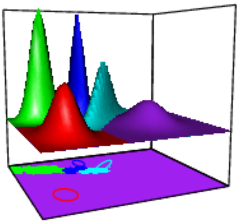
MLE for Classification



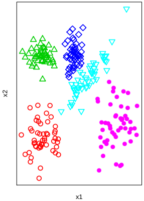
features: $\mathbf{x} \in \mathbb{R}^d$

class: $y \in \{y_1, \ldots, y_K\}$

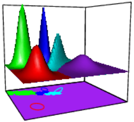
$$k_{\text{MAP}} = \arg\max_k Pr(y_k|\mathbf{x})$$
$$= \arg\max_k Pr(\mathbf{x}|y_k)Pr(y_k)$$



Assumption: Class Independence

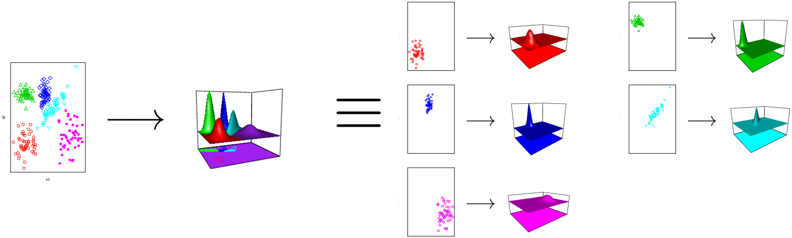


$\longrightarrow$



samples from class i do not influence estimate for class j, $i \neq j$

Assumption: Class Independence



- distribution of $\mathbf{x}$ for class y_k is the likelihood $Pr(\mathbf{x}|\theta_k)$
- in the following, we drop the class index k and write $Pr(\mathbf{x}|\theta)$
- also we call $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ the set of data point belonging to a single class y_k

ML estimation of Gaussian mean

$$X \sim \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \text{ with } \theta = \{\mu, \sigma^2\}$$

Log-likelihood of data (i.i.d. samples):

$$\log Pr(\mathcal{D}|\theta) = \sum_{n=1}^{N} \log \mathcal{N}(x_n|\mu, \sigma^2)$$

ML estimation of Gaussian mean

$$X \sim \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \text{ with } \theta = \{\mu, \sigma^2\}$$

Log-likelihood of data (i.i.d. samples):

$$\log Pr(\mathcal{D}|\theta) = \sum_{n=1}^{N} \log \mathcal{N}(x_n|\mu, \sigma^2) = -N \log\left(\sqrt{2\pi\sigma^2}\right) - \sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}$$

ML estimation of Gaussian mean

$$X \sim \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \text{ with } \theta = \{\mu,\sigma^2\}$$

Log-likelihood of data (i.i.d. samples):

$$\log Pr(\mathcal{D}|\theta) = \sum_{n=1}^{N} \log \mathcal{N}(x_n|\mu,\sigma^2) = -N \log\left(\sqrt{2\pi\sigma^2}\right) - \sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}$$

$$0 = \frac{d \log Pr(\mathcal{D}|\theta)}{d\mu}$$

ML estimation of Gaussian mean

$$X \sim \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \text{ with } \theta = \{\mu, \sigma^2\}$$

Log-likelihood of data (i.i.d. samples):

$$\log Pr(\mathcal{D}|\theta) = \sum_{n=1}^{N} \log \mathcal{N}(x_n|\mu, \sigma^2) = -N \log\left(\sqrt{2\pi\sigma^2}\right) - \sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}$$

$$0 = \frac{d \log Pr(\mathcal{D}|\theta)}{d\mu} = \sum_{n=1}^{N} \frac{(x_n-\mu)}{2\sigma^2}$$

ML estimation of Gaussian mean

$$X \sim \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \text{ with } \theta = \{\mu,\sigma^2\}$$

Log-likelihood of data (i.i.d. samples):

$$\log Pr(\mathcal{D}|\theta) = \sum_{n=1}^{N} \log \mathcal{N}(x_n|\mu,\sigma^2) = -N\log\left(\sqrt{2\pi\sigma^2}\right) - \sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}$$

$$0 = \frac{d \log Pr(\mathcal{D}|\theta)}{d\mu} = \sum_{n=1}^{N} \frac{(x_n-\mu)}{2\sigma^2} = \frac{\sum_{n=1}^{N} x_n - N\mu}{2\sigma^2} \iff$$

ML estimation of Gaussian mean

$$X \sim \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \text{ with } \theta = \{\mu,\sigma^2\}$$

Log-likelihood of data (i.i.d. samples):

$$\log Pr(\mathcal{D}|\theta) = \sum_{n=1}^{N} \log \mathcal{N}(x_n|\mu,\sigma^2) = -N \log\left(\sqrt{2\pi\sigma^2}\right) - \sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}$$

$$0 = \frac{d \log Pr(\mathcal{D}|\theta)}{d\mu} = \sum_{n=1}^{N} \frac{(x_n-\mu)}{2\sigma^2} = \frac{\sum_{n=1}^{N} x_n - N\mu}{2\sigma^2} \iff$$

$$\mu_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

ML estimation of Gaussian parameters

$$\begin{aligned}
\mu_{\mathsf{ML}} &= \frac{1}{N}\sum_{n=1}^{N} x_n \\
\sigma^2_{\mathsf{ML}} &= \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{\mathsf{ML}})^2
\end{aligned}$$

ML estimation of Gaussian parameters

$$\mu_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma^2_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{\text{ML}})^2$$

- This is the same result as minimizing the sum of square errors!
- but now out assumptions are explicit (i.e., how the data is distributed)
- This estimate of the variance is *biased*, i.e., $\mathbb{E}[\sigma^2_{\text{ML}}] - \sigma^2 \neq 0$. The unbiased ML estimate is

$$\sigma'^2_{\text{ML}} = \frac{1}{N-1}\sum_{n=1}^{N} (x_n - \mu_{\text{ML}})^2$$

MLE with Discrete Variables

Will I go and play orienteering given the forecast?

$$x \in \{\text{sunny}, \text{overcast}, \text{rainy}\}$$
$$y \in \{\text{yes}, \text{no}\}$$

$$X \sim ?$$
$$Y \sim ?$$
$$X|Y \sim ?$$
$$Y|X \sim ?$$

Training data

n example	x_n outlook	y_n play	n example	x_n outlook	y_n play
1	sunny	no	8	sunny	no
2	sunny	no	9	sunny	yes
3	overcast	yes	10	rainy	yes
4	rainy	yes	11	sunny	yes
5	rainy	yes	12	overcast	yes
6	rainy	no	13	overcast	yes
7	overcast	yes	14	rainy	no

MLE with Discrete Variables

Will I go and play orienteering given the forecast?

$$
\begin{aligned}
x &\in \{\text{sunny}, \text{overcast}, \text{rainy}\} \\
y &\in \{\text{yes}, \text{no}\}
\end{aligned}
$$

$$
\begin{aligned}
X &\sim \text{Cat}(\lambda_1, \lambda_2, \lambda_3) \\
Y &\sim ? \\
X|Y &\sim ? \\
Y|X &\sim ?
\end{aligned}
$$

Training data

n example	x_n outlook	y_n play	n example	x_n outlook	y_n play
1	sunny	no	8	sunny	no
2	sunny	no	9	sunny	yes
3	overcast	yes	10	rainy	yes
4	rainy	yes	11	sunny	yes
5	rainy	yes	12	overcast	yes
6	rainy	no	13	overcast	yes
7	overcast	yes	14	rainy	no

MLE with Discrete Variables

Will I go and play orienteering given the forecast?

$$\begin{aligned} x &\in \{\text{sunny}, \text{overcast}, \text{rainy}\} \\ y &\in \{\text{yes}, \text{no}\} \end{aligned}$$

$$\begin{aligned} X &\sim \text{Cat}(\lambda_1, \lambda_2, \lambda_3) \\ Y &\sim \text{Bernoulli}(\lambda) \\ X|Y &\sim ? \\ Y|X &\sim ? \end{aligned}$$

Training data

n example	x_n outlook	y_n play	n example	x_n outlook	y_n play
1	sunny	no	8	sunny	no
2	sunny	no	9	sunny	yes
3	overcast	yes	10	rainy	yes
4	rainy	yes	11	sunny	yes
5	rainy	yes	12	overcast	yes
6	rainy	no	13	overcast	yes
7	overcast	yes	14	rainy	no

MLE with Discrete Variables

Will I go and play orienteering given the forecast?

$$x \in \{\text{sunny}, \text{overcast}, \text{rainy}\}$$
$$y \in \{\text{yes}, \text{no}\}$$

$$X \sim \text{Cat}(\lambda_1, \lambda_2, \lambda_3)$$
$$Y \sim \text{Bernoulli}(\lambda)$$
$$X|Y \sim \text{Cat}(\lambda'_1, \lambda'_2, \lambda'_3)$$
$$Y|X \sim \text{Bernoulli}(\lambda')$$

Training data

n example	x_n outlook	y_n play	n example	x_n outlook	y_n play
1	sunny	no	8	sunny	no
2	sunny	no	9	sunny	yes
3	overcast	yes	10	rainy	yes
4	rainy	yes	11	sunny	yes
5	rainy	yes	12	overcast	yes
6	rainy	no	13	overcast	yes
7	overcast	yes	14	rainy	no

MLE: Bernoulli

$$Pr(y) = \begin{cases} \lambda & \text{if } y = \text{yes} \\ 1-\lambda & \text{if } y = \text{no} \end{cases}$$

1. compute (log) likelihood of the data $P(\mathcal{D}|\lambda)$
2. find λ_{ML} that optimizes $P(\mathcal{D}|\lambda)$

n example	x_n outlook	y_n play	n example	x_n outlook	y_n play
1	sunny	no	8	sunny	no
2	sunny	no	9	sunny	yes
3	overcast	yes	10	rainy	yes
4	rainy	yes	11	sunny	yes
5	rainy	yes	12	overcast	yes
6	rainy	no	13	overcast	yes
7	overcast	yes	14	rainy	no

MLE: Bernoulli

$$Pr(y) = \begin{cases} \lambda & \text{if } y = \text{yes} \\ 1-\lambda & \text{if } y = \text{no} \end{cases}$$

Likelihood of the data
(n=number of yes in $\mathcal{D}$, N=number of examples):

$$\begin{aligned} Pr(\mathcal{D}|\lambda) &= \prod_n Pr(y_n|\lambda) = \prod_{n \text{ s.t. } y = \text{yes}} \lambda \prod_{n \text{ s.t. } y = \text{no}} (1-\lambda) \\ &= \lambda^n (1-\lambda)^{N-n} \end{aligned}$$

MLE: Bernoulli

$$Pr(y) = \begin{cases} \lambda & \text{if } y = \text{yes} \\ 1-\lambda & \text{if } y = \text{no} \end{cases}$$

Likelihood of the data
(n=number of yes in $\mathcal{D}$, N=number of examples):

$$\begin{aligned} Pr(\mathcal{D}|\lambda) &= \prod_n Pr(y_n|\lambda) = \prod_{n \text{ s.t. } y = \text{yes}} \lambda \prod_{n \text{ s.t. } y = \text{no}} (1-\lambda) \\ &= \lambda^n (1-\lambda)^{N-n} \\ \log Pr(\mathcal{D}|\lambda) &= n \log \lambda + (N-n) \log(1-\lambda) \end{aligned}$$

MLE: Bernoulli

$$Pr(y) = \begin{cases} \lambda & \text{if } y = \text{yes} \\ 1-\lambda & \text{if } y = \text{no} \end{cases}$$

Likelihood of the data
(n=number of yes in $\mathcal{D}$, N=number of examples):

$$\begin{aligned} Pr(\mathcal{D}|\lambda) &= \prod_n Pr(y_n|\lambda) = \prod_{n \text{ s.t. } y = \text{yes}} \lambda \prod_{n \text{ s.t. } y = \text{no}} (1-\lambda) \\ &= \lambda^n (1-\lambda)^{N-n} \\ \log Pr(\mathcal{D}|\lambda) &= n \log \lambda + (N-n) \log(1-\lambda) \\ \frac{d}{d\lambda} \log Pr(\mathcal{D}|\lambda) &= \frac{n - N\lambda}{\lambda(1-\lambda)} = 0 \end{aligned}$$

MLE: Bernoulli

$$Pr(y) = \begin{cases} \lambda & \text{if } y = \text{yes} \\ 1-\lambda & \text{if } y = \text{no} \end{cases}$$

Likelihood of the data
(n=number of yes in $\mathcal{D}$, N=number of examples):

$$\begin{aligned} Pr(\mathcal{D}|\lambda) &= \prod_n Pr(y_n|\lambda) = \prod_{n \text{ s.t. } y=\text{yes}} \lambda \prod_{n \text{ s.t. } y=\text{no}} (1-\lambda) \\ &= \lambda^n (1-\lambda)^{N-n} \\ \log Pr(\mathcal{D}|\lambda) &= n \log \lambda + (N-n)\log(1-\lambda) \\ \frac{d}{d\lambda} \log Pr(\mathcal{D}|\lambda) &= \frac{n - N\lambda}{\lambda(1-\lambda)} = 0 \iff \lambda_{\text{ML}} = \frac{n}{N} \end{aligned}$$

MLE Example: Discrete Variables

Will I go and play orienteering given the forecast?

$$x \in \{\text{sunny}, \text{overcast}, \text{rainy}\}$$
$$y \in \{\text{yes}, \text{no}\}$$

Training data

n example	x_n outlook	y_n play	n example	x_n outlook	y_n play
1	sunny	no	8	sunny	no
2	sunny	no	9	sunny	yes
3	overcast	yes	10	rainy	yes
4	rainy	yes	11	sunny	yes
5	rainy	yes	12	overcast	yes
6	rainy	no	13	overcast	yes
7	overcast	yes	14	rainy	no

$$Y \sim \text{Bernoulli}(\lambda)$$
$$\lambda_{\text{ML}} = \frac{9}{14}$$

MLE: Categorical

Similar derivation:

$$\lambda_{k,\text{ML}} = \frac{n_k}{N}$$

where n_k is the number of examples of the kth category

$$X \sim \text{Cat}(\lambda_{\text{sunny}}, \lambda_{\text{overcast}}, \lambda_{\text{rainy}})$$

$$\lambda_{\text{ML}} = \left\{ \frac{5}{14}, \frac{4}{14}, \frac{5}{14} \right\}$$

Training data

n example	x_n outlook	y_n play	n example	x_n outlook	y_n play
1	sunny	no	8	sunny	no
2	sunny	no	9	sunny	yes
3	overcast	yes	10	rainy	yes
4	rainy	yes	11	sunny	yes
5	rainy	yes	12	overcast	yes
6	rainy	no	13	overcast	yes
7	overcast	yes	14	rainy	no

MLE: Categorical

Similar derivation:

$$\lambda_{k,\text{ML}} = \frac{n_k}{N}$$

where n_k is the number of examples of the kth category

$$X \sim \text{Cat}(\lambda_{\text{sunny}}, \lambda_{\text{overcast}}, \lambda_{\text{rainy}})$$

$$\lambda_{\text{ML}} = \left\{ \frac{5}{14}, \frac{4}{14}, \frac{5}{14} \right\}$$

$$X|Y \sim \text{Cat}(\lambda'_1, \ldots, \lambda'_k)$$

Training data

n example	x_n outlook	y_n play	n example	x_n outlook	y_n play
1	sunny	no	8	sunny	no
2	sunny	no	9	sunny	yes
3	overcast	yes	10	rainy	yes
4	rainy	yes	11	sunny	yes
5	rainy	yes	12	overcast	yes
6	rainy	no	13	overcast	yes
7	overcast	yes	14	rainy	no

MLE: Categorical

Similar derivation:

$$\lambda_{k,\text{ML}} = \frac{n_k}{N}$$

where n_k is the number of examples of the kth category

$$X \sim \text{Cat}(\lambda_{\text{sunny}}, \lambda_{\text{overcast}}, \lambda_{\text{rainy}})$$

$$\lambda_{\text{ML}} = \left\{ \frac{5}{14}, \frac{4}{14}, \frac{5}{14} \right\}$$

$$X|Y \sim \text{Cat}(\lambda'_1, \ldots, \lambda'_k)$$

$$\lambda'_{\text{ML}}(\text{yes}) = \left\{ \frac{2}{9}, \frac{4}{9}, \frac{3}{9} \right\}$$

$$\lambda'_{\text{ML}}(\text{no}) = \left\{ \frac{3}{5}, 0, \frac{2}{5} \right\}$$

Training data

n example	x_n outlook	y_n play	n example	x_n outlook	y_n play
1	sunny	no	8	sunny	no
2	sunny	no	9	sunny	yes
3	overcast	yes	10	rainy	yes
4	rainy	yes	11	sunny	yes
5	rainy	yes	12	overcast	yes
6	rainy	no	13	overcast	yes
7	overcast	yes	14	rainy	no

But ... will I play orienteering given a rainy outlook?

$$
\begin{aligned}
Pr(y = \text{yes}|\text{outlook=rainy}) &= \frac{Pr(\text{outlook=rainy}|y = \text{yes})Pr(y = \text{yes})}{Pr(\text{outlook=rainy})} \\
&= \frac{\frac{3}{9}\frac{9}{14}}{\frac{5}{14}} = \frac{3}{5}
\end{aligned}
$$

But . . . will I play orienteering given a rainy outlook?

$$
\begin{aligned}
Pr(y = \text{yes}|\text{outlook=rainy}) &= \frac{Pr(\text{outlook=rainy}|y = \text{yes})Pr(y = \text{yes})}{Pr(\text{outlook=rainy})} \\
&= \frac{\frac{3}{9}\frac{9}{14}}{\frac{5}{14}} = \frac{3}{5} \\
Pr(y = \text{no}|\text{outlook=rainy}) &= \frac{Pr(\text{outlook=rainy}|y = \text{no})Pr(y = \text{no})}{Pr(\text{outlook=rainy})} \\
&= \frac{\frac{2}{5}\frac{5}{14}}{\frac{5}{14}} = \frac{2}{5}
\end{aligned}
$$

Then

$$
y_{\text{MAP}} = \arg\max_y Pr(y|\text{outlook=rainy}) = \text{yes } (3/5 > 2/5)
$$

But . . . will I play orienteering given a rainy outlook?

$$
\begin{aligned}
Pr(y = \text{yes}|\text{outlook=rainy}) &= \frac{Pr(\text{outlook=rainy}|y = \text{yes})Pr(y = \text{yes})}{Pr(\text{outlook=rainy})} \\
&= \frac{\frac{3}{9}\frac{9}{14}}{\frac{5}{14}} = \frac{3}{5} \\
Pr(y = \text{no}|\text{outlook=rainy}) &= \frac{Pr(\text{outlook=rainy}|y = \text{no})Pr(y = \text{no})}{Pr(\text{outlook=rainy})} \\
&= \frac{\frac{2}{5}\frac{5}{14}}{\frac{5}{14}} = \frac{2}{5}
\end{aligned}
$$

Then

$$
\begin{aligned}
y_{\text{MAP}} &= \arg\max_y Pr(y|\text{outlook=rainy}) = \text{yes } (3/5 > 2/5) \\
y_{\text{ML}} &= \arg\max_y Pr(\text{outlook=rainy}|y) = \text{no } (2/5 > 3/9)
\end{aligned}
$$

Source of confusion

Maximum a Posteriori (MAP) and Maximum Likelihood (ML) classification are *different*:

$$
\begin{aligned}
y_{\mathsf{MAP}} &= \arg\max_y P(y|x, \theta_{\mathsf{ML}}) \\
y_{\mathsf{ML}} &= \arg\max_y P(x|y, \theta_{\mathsf{ML}})
\end{aligned}
$$

even with parameters θ estimated with the ML *optimality criterion*:

$$
\theta_{\mathsf{ML}} = \arg\max_\theta P(D|y,\theta) = \arg\max_\theta \prod_n P(x_n|y_n,\theta)
$$

NB: ML *parameter* estimation is not ML regression/classification.

Outline

Problem: Curse of Dimensionality

n example	$\mathbf{x}_n$ outlook	temperature	humidity	windy	y_n play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

difficult to model $Pr(\text{outlook}, \text{temperature}, \text{humidity}, \text{windy}|\text{play})$

Problem: Curse of Dimensionality

- Volume of feature space exponential in number of features.
- More features $\implies$ potential for better description of the objects but ...
- ... $\implies$ need more and more data to model $Pr(x, y)$ well

Problem: Curse of Dimensionality

- Volume of feature space exponential in number of features.
- More features $\implies$ potential for better description of the objects but ...
- ... $\implies$ need more and more data to model $Pr(x, y)$ well

Approximation: **Naïve Bayes classifier**

- All features (dimensions) regarded as conditionally independent.
- Instead of modelling one D-dimensional distribution: $Pr(\text{outlook, temperature, humidity, windy}|\text{play})$ model D one-dimensional distributions: $Pr(\text{outlook}|\text{play})$, $Pr(\text{temperature}|\text{play})$, $Pr(\text{humidity}|\text{play})$, $Pr(\text{windy}|\text{play})$

Naïve Bayes Classifier

- $\mathbf{x}$ is a vector $(x_1, \ldots, x_D)$ of attribute or feature values.
- Let $\mathcal{Y} = \{1, 2, \ldots, K\}$ be the set of possible classes.
- MAP classification is

$$y_{\text{MAP}} = \arg\max_{y \in \mathcal{Y}} Pr(y \,|\, x_1, \ldots, x_D) = \arg\max_{y \in \mathcal{Y}} \frac{Pr(x_1, \ldots, x_D \,|\, y)\, Pr(y)}{Pr(x_1, \ldots, x_D)}$$
$$= \arg\max_{y \in \mathcal{Y}} Pr(x_1, \ldots, x_D \,|\, y)\, Pr(y)$$

Naïve Bayes Classifier

- $\mathbf{x}$ is a vector $(x_1, \ldots, x_D)$ of attribute or feature values.
- Let $\mathcal{Y} = \{1, 2, \ldots, K\}$ be the set of possible classes.
- MAP classification is

$$\begin{aligned} y_{\text{MAP}} &= \arg\max_{y \in \mathcal{Y}} Pr(y \mid x_1, \ldots, x_D) = \arg\max_{y \in \mathcal{Y}} \frac{Pr(x_1, \ldots, x_D \mid y)\, Pr(y)}{Pr(x_1, \ldots, x_D)} \\ &= \arg\max_{y \in \mathcal{Y}} Pr(x_1, \ldots, x_D \mid y)\, Pr(y) \end{aligned}$$

- Naïve Bayes assumption:
  $Pr(x_1, \ldots, x_D \mid y) = \prod_{d=1}^{D} Pr(x_d \mid y)$

Naïve Bayes Classifier

- $\mathbf{x}$ is a vector $(x_1, \ldots, x_D)$ of attribute or feature values.
- Let $\mathcal{Y} = \{1, 2, \ldots, K\}$ be the set of possible classes.
- MAP classification is

$$y_{\text{MAP}} = \arg\max_{y \in \mathcal{Y}} Pr(y \,|\, x_1, \ldots, x_D) = \arg\max_{y \in \mathcal{Y}} \frac{Pr(x_1, \ldots, x_D \,|\, y)\, Pr(y)}{Pr(x_1, \ldots, x_D)}$$
$$= \arg\max_{y \in \mathcal{Y}} Pr(x_1, \ldots, x_D \,|\, y)\, Pr(y)$$

- Naïve Bayes assumption:
  $Pr(x_1, \ldots, x_D \,|\, y) = \prod_{d=1}^{D} Pr(x_d \,|\, y)$
- MAP classification with Naïve Bayes:

$$y_{\text{MAP}} = \arg\max_{y \in \mathcal{Y}} Pr(y) \prod_{d=1}^{D} Pr(x_d \,|\, y)$$

Naïve Bayes Classifier

$$y_{\text{MAP}} = \arg\max_{y \in \mathcal{Y}} Pr(y) \prod_{d=1}^{D} Pr(x_d \,|\, y)$$

Naïve Bayes is one of the most common learning methods.
When to use:

- Moderate or large training set available.
- Feature dimensions are conditionally independent given class (or at least reasonably independent, still works with a little dependence).

Successful applications:

- Medical diagnoses (symptoms independent)
- Classification of text documents (words independent)

Example: Play Orienteering?

Question: Will I go and play orienteering given the forecast?

My measurements:

- **outlook** $\in$ {sunny, overcast, rainy},
- **temperature** $\in$ {hot, mild, cool},
- **humidity** $\in$ {high, normal},
- **windy** $\in$ {false, true}.

Possible decisions: $y \in$ {yes, no}

Example: Play Orienteering?

What I did in the past:

n	$\mathbf{x}_n$				y_n
example	outlook	temperature	humidity	windy	play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

xample: Play Orienteering?

Counts of when I played orienteering (did not play)

Outlook			**Temperature**			**Humidity**		**Windy**	
sunny	overcast	rain	hot	mild	cool	high	normal	false	true
2 (3)	4 (0)	3 (2)	2 (2)	4 (2)	3 (1)	3 (4)	6 (1)	6 (2)	3 (3)

xample: Play Orienteering?

Counts of when I played orienteering (did not play)

Outlook			**Temperature**			**Humidity**		**Windy**	
sunny	overcast	rain	hot	mild	cool	high	normal	false	true
2 (3)	4 (0)	3 (2)	2 (2)	4 (2)	3 (1)	3 (4)	6 (1)	6 (2)	3 (3)

Prior of whether I played orienteering or not

Counts:

Play	
yes	no
9	5

Prior Probabilities:

Play	
yes	no
$\frac{9}{14}$	$\frac{5}{14}$

xample: Play Orienteering?

Counts of when I played orienteering (did not play)

Outlook			Temperature			Humidity		Windy	
sunny	overcast	rain	hot	mild	cool	high	normal	false	true
2 (3)	4 (0)	3 (2)	2 (2)	4 (2)	3 (1)	3 (4)	6 (1)	6 (2)	3 (3)

Prior of whether I played orienteering or not

Counts:

Play	
yes	no
9	5

Prior Probabilities:

Play	
yes	no
$\frac{9}{14}$	$\frac{5}{14}$

Likelihood of attribute when orienteering played $Pr(x_i \,|\, \text{y=yes})(Pr(x_i \,|\, \text{y=no}))$

Outlook			Temperature			Humidity		Windy	
sunny	overcast	rain	hot	mild	cool	high	normal	false	true
$\frac{2}{9}$ $(\frac{3}{5})$	$\frac{4}{9}$ $(\frac{0}{5})$	$\frac{3}{9}$ $(\frac{2}{5})$	$\frac{2}{9}$ $(\frac{2}{5})$	$\frac{4}{9}$ $(\frac{2}{5})$	$\frac{3}{9}$ $(\frac{1}{5})$	$\frac{3}{9}$ $(\frac{4}{5})$	$\frac{6}{9}$ $(\frac{1}{5})$	$\frac{6}{9}$ $(\frac{2}{5})$	$\frac{3}{9}$ $(\frac{3}{5})$

Example: Play Orienteering?

Inference: Use the learnt model to classify a new instance.

New instance:

$$\mathbf{x} = (\text{sunny, cool, high, true})$$

Apply Naïve Bayes Classifier:

$$y_{\text{MAP}} = \arg \max_{y \,\in\, \{\text{yes, no}\}} Pr(y) \prod_{i=1}^{4} Pr(x_i \,|\, y)$$

$$P(\text{yes})\, P(\text{sunny} \,|\, \text{yes})\, P(\text{cool} \,|\, \text{yes})\, P(\text{high} \,|\, \text{yes})\, P(\text{true} \,|\, \text{yes}) = \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = .005$$

$$P(\text{no})\, P(\text{sunny} \,|\, \text{no})\, P(\text{cool} \,|\, \text{no})\, P(\text{high} \,|\, \text{no})\, P(\text{true} \,|\, \text{no}) = \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = .021$$

Example: Play Orienteering?

Inference: Use the learnt model to classify a new instance.

New instance:

$$\mathbf{x} = (\text{sunny, cool, high, true})$$

Apply Naïve Bayes Classifier:

$$y_{\text{MAP}} = \arg \max_{y \,\in\, \{\text{yes, no}\}} Pr(y) \prod_{i=1}^{4} Pr(x_i \,|\, y)$$

$$P(\text{yes})\ P(\text{sunny} \,|\, \text{yes})\ P(\text{cool} \,|\, \text{yes})\ P(\text{high} \,|\, \text{yes})\ P(\text{true} \,|\, \text{yes}) = \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = .005$$

$$P(\text{no})\ P(\text{sunny} \,|\, \text{no})\ P(\text{cool} \,|\, \text{no})\ P(\text{high} \,|\, \text{no})\ P(\text{true} \,|\, \text{no}) = \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = .021$$

$\Longrightarrow \ y_{\text{MAP}} = \text{no}$

Naïve Bayes: Independence Violation

- Conditional independence assumption:

$$Pr(x_1, x_2, \ldots, x_D \,|\, y) = \prod_{d=1}^{D} Pr(x_d \,|\, y)$$

  often violated – but it works surprisingly well anyway!

- Since dependencies ignored, naïve Bayes posteriors often unrealistically close to 0 or 1.

  Different attributes say the same thing to a higher degree than we expect as they are correlated in reality.

Naïve Bayes: Estimating Probabilities

- **Problem:** What if none of the training instances with target value y have attribute x_i? Then

$$Pr(x_i \,|\, y) = 0 \quad \Longrightarrow \quad Pr(y) \prod_{i=1}^{D} Pr(x_i \,|\, y) = 0$$

Naïve Bayes: Estimating Probabilities

- **Problem:** What if none of the training instances with target value y have attribute x_i? Then

$$Pr(x_i \,|\, y) = 0 \quad \implies \quad Pr(y)\prod_{i=1}^{D} Pr(x_i \,|\, y) = 0$$

- **Simple solution:** add pseudocounts to all counts so that no count is zero

Naïve Bayes: Estimating Probabilities

- **Problem:** What if none of the training instances with target value y have attribute x_i? Then

$$Pr(x_i \,|\, y) = 0 \quad \implies \quad Pr(y)\prod_{i=1}^{D} Pr(x_i \,|\, y) = 0$$

- **Simple solution:** add pseudocounts to all counts so that no count is zero
- This is a form of regularization or smoothing

Logistic Regression

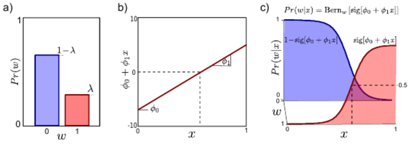


Figure from Prince (Ch. 9)

- Binary classification problem: $y \in \{0, 1\}$ treated as a regression problem: $\mathbf{x} \to \lambda$ (Bernoulli param.)

$$
\begin{aligned}
Y|\mathbf{X} &\sim \text{Bernoulli}(\lambda(\mathbf{x})) \\
Pr(y|\mathbf{x}) &= \lambda(\mathbf{x})^y (1 - \lambda(\mathbf{x}))^{(1-y)} \\
\lambda(\mathbf{x}) &= \text{sigmoid}(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}
\end{aligned}
$$

Logistic Regression

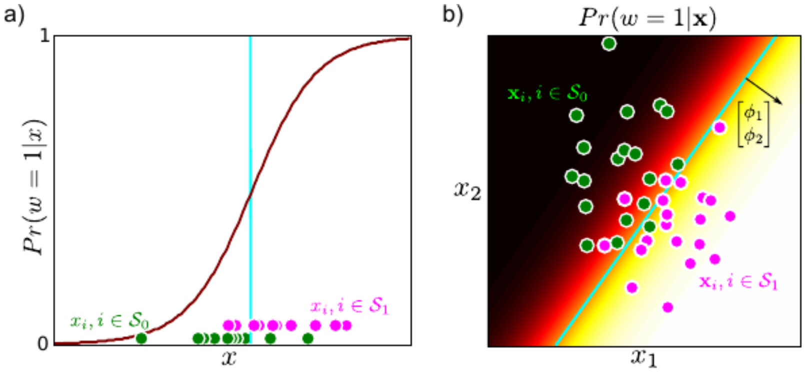
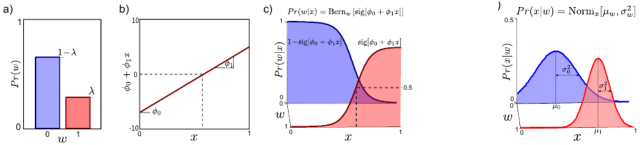


Figure from Prince (Ch. 9)

Logistic Regression vs Gaussian Classifier



Figures from Prince (Ch. 9)

Different learning:

- Gaussians: generative model, optimize $Pr(\mathbf{x}|y_0)$ and $Pr(\mathbf{x}|y_1)$
- Logistic Regression: discriminative model, optimize $Pr(y_1|\mathbf{x})$

Logistic Regression: MLE

Learning: maximize $Pr(y|\mathbf{x})$ (discriminative method)

$$
\begin{aligned}
Pr(y|\mathbf{x},\mathbf{w}) &= \prod_{i=1}^{N} \lambda(\mathbf{x}_i)^{y_i}(1-\lambda(\mathbf{x}_i))^{(1-y_i)} \\
\log Pr(y|\mathbf{x},\mathbf{w}) &= \sum_{i=1}^{N} [y_i \log \lambda(\mathbf{x}_i) + (1-y_i)\log(1-\lambda(\mathbf{x}_i))] \\
&= \sum_{i=1}^{N} \left[y_i \log \mathrm{sig}(\mathbf{w}^T\mathbf{x}_i) + (1-y_i)\log\left(1-\mathrm{sig}(\mathbf{w}^T\mathbf{x}_i)\right)\right]
\end{aligned}
$$

Logistic Regression: MLE

Learning: maximize $Pr(y|\mathbf{x})$ (discriminative method)

$$
\begin{aligned}
Pr(y|\mathbf{x},\mathbf{w}) &= \prod_{i=1}^{N} \lambda(\mathbf{x}_i)^{y_i}(1-\lambda(\mathbf{x}_i))^{(1-y_i)} \\
\log Pr(y|\mathbf{x},\mathbf{w}) &= \sum_{i=1}^{N} [y_i \log \lambda(\mathbf{x}_i) + (1-y_i)\log(1-\lambda(\mathbf{x}_i))] \\
&= \sum_{i=1}^{N} \left[y_i \log \mathrm{sig}(\mathbf{w}^T\mathbf{x}_i) + (1-y_i)\log\left(1-\mathrm{sig}(\mathbf{w}^T\mathbf{x}_i)\right)\right]
\end{aligned}
$$

Optimize by setting: no close form solution! Use gradient descent

$$
\frac{d}{d\mathbf{w}} \log Pr(y|\mathbf{x},\mathbf{w}) = \sum_{i=1}^{N} \left(y_i - \mathrm{sig}(\mathbf{w}^T\mathbf{x}_i)\right)\mathbf{x}_i = 0
$$

Hints: derivatives of sigmoid

$$\mathsf{sig}(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$

$$\frac{d}{d\mathbf{w}} \mathrm{sig}(\mathbf{w}^T\mathbf{x}) = \mathrm{sig}(\mathbf{w}^T\mathbf{x}) \left(1 - \mathrm{sig}(\mathbf{w}^T\mathbf{x})\right) \mathbf{x}$$

$$\frac{d}{d\mathbf{w}} \log\left(\mathrm{sig}(\mathbf{w}^T\mathbf{x})\right) = \frac{\mathrm{sig}(\mathbf{w}^T\mathbf{x}) \left(1 - \mathrm{sig}(\mathbf{w}^T\mathbf{x})\right)}{\mathrm{sig}(\mathbf{w}^T\mathbf{x})}\mathbf{x} = \left(1 - \mathrm{sig}(\mathbf{w}^T\mathbf{x})\right) \mathbf{x}$$

$$\frac{d}{d\mathbf{w}} \log\left(1 - \mathrm{sig}(\mathbf{w}^T\mathbf{x})\right) = \frac{-\mathrm{sig}(\mathbf{w}^T\mathbf{x}) \left(1 - \mathrm{sig}(\mathbf{w}^T\mathbf{x})\right)}{1 - \mathrm{sig}(\mathbf{w}^T\mathbf{x})}\mathbf{x} = -\mathrm{sig}(\mathbf{w}^T\mathbf{x})\ \mathbf{x}$$

Logistic Regression vs Conditional Gaussian

Number of parameters (D dimensions, 2 classes):

Gaussian distributions (equal priors)

$2 \times D$	(mean vectors)
$D(D+1)/2$	(shared covariance)
$D(D+5)/2$	(total, quadratic in D)

Logistic Regression

D (weights)

Training:

Gaussian distributions

- closed form solution
- generative model

Logistic Regression

- gradient descent
- discriminative model

Summary

1. Introduction
   - Probabilistic Classification and Regression
   - Discriminative vs Generative Models
   - Parametric vs Non-parametric Inference
2. Maximum Likelihood (ML) Estimation
   - Regression
   - Classification
3. Special Cases
   - Naïve Bayes Classifier
   - Logistic Regression

Check your understanding!

Consider a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$ of N independent and identically distributed observations where each $\mathbf{x}_n$ is a p-dimensional real vector. Assume the random variable Y_n is distributed Laplacian with a mean $\boldsymbol{\beta}^T \mathbf{x}_n$ and known scale parameter $b > 0$. In other words, $Y_n | \mathbf{x}_n, \boldsymbol{\beta} \sim \mathcal{L}(\boldsymbol{\beta}^T \mathbf{x}_n, b)$. Define the a priori distribution of the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ *multivariate* Gaussian with parameters mean $\mathbf{0}$ and variance $\sigma^2 \mathbf{I}$.

Check your understanding!

1. Derive the maximum likelihood (ML) estimate of $\boldsymbol{\beta}$.
2. Which of the following statements is not true? (There may be more than one.)
    1. As b increases and N remains constant, the ML estimate of $\boldsymbol{\beta}$ becomes poorer.
    2. As N increases and b remains constant, the ML estimate of $\boldsymbol{\beta}$ becomes poorer.
    3. As b decreases and N remains constant, the ML estimate of $\boldsymbol{\beta}$ becomes poorer.
    4. As N decreases and b remains constant, the ML estimate of $\boldsymbol{\beta}$ becomes poorer.