Maximum Likelihood & Naïve Bayes

James McInerney

Adapted from slides by Nakul Verma

Supervised Machine Learning

Statistical modeling approach:

Labeled training data
(*n* examples from data) → Learning Algorithm → 'classifier' $\hat{f}$

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots (\vec{x}_n, y_n)$$

drawn **independently** from
a fixed underlying distribution
(also called the *i.i.d.* assumption)

select $\hat{f}$ from…?
from a pool of ***models*** $\mathcal{F}$
that **maximizes**
label agreement of the
training data

How to select $\hat{f} \in \mathcal{F}$?

- Maximum likelihood (best fits the data)
- Maximum a posteriori (best fits the data but incorporates prior assumptions)
- Optimization of 'loss' criterion (best discriminates the labels)
- …

Maximum Likelihood Estimation (MLE)

Given some data $\vec{x}_1, \vec{x}_2, \ldots \vec{x}_n \in \mathcal{X}$ i.i.d. *(Let's forget about the labels for now)*

Say we have a model class $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$

ie, each model *p* can be described by a set of parameters θ

find the parameter settings θ that **best fits** the data.

If each model *p*, is a **probability model** then we can find the best fitting probability model via the ***likelihood estimation***!

Likelihood $\mathcal{L}(\theta|X) := P(X|\theta) = P(\vec{x}_1, \ldots, \vec{x}_n|\theta) \overset{\textit{i.i.d.}}{=} \prod_{i=1}^{n} P(\vec{x}_i|\theta) = \prod_{i=1}^{n} p_\theta(\vec{x}_i)$

Interpretation: How **probable** (or how likely) is the data given the model p_θ?

Parameter setting θ that maximizes $\mathcal{L}(\theta|X)$

$$\arg\max_\theta \mathcal{L}(\theta|X) = \arg\max_\theta \prod_{i=1}^{n} p_\theta(\vec{x}_i)$$

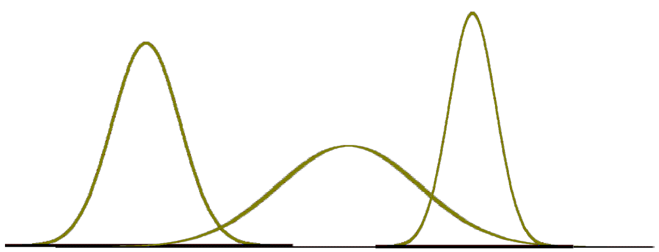
MLE Example

Fitting a model to heights of females

Height data (in inches): 60, 62, 53, 58, … $\in \mathbf{R}$

$x_1, x_2, \ldots x_n \in \mathcal{X}$

Model class: Gaussian models in **R**



$$p_\theta(x) = p_{\{\mu,\sigma^2\}}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

μ = mean parameter

σ^2 = variance parameter > 0

So, what is the MLE for the given data X ?

MLE Example (contd.)

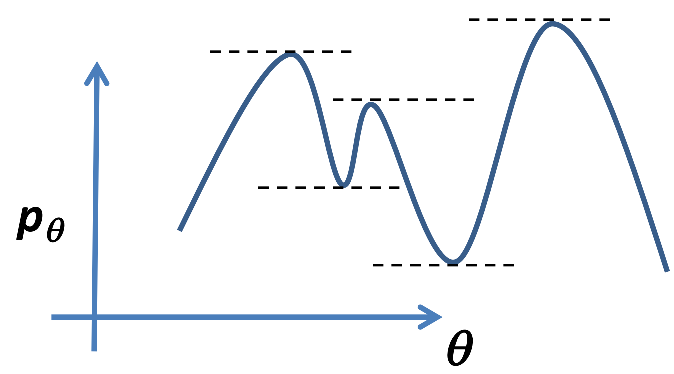
Height data (in inches): $x_1, x_2, \dots x_n \in \mathcal{X}$ = **R**

Model class: Gaussian models in **R** $\quad p_{\{\mu,\sigma^2\}}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$

MLE: $\quad \arg\max_{\theta} \mathcal{L}(\theta|X) = \arg\max_{\mu,\sigma^2} \prod_{i=1}^{n} p_{\{\mu,\sigma^2\}}(x_i)$ *Good luck!*

Trick #1: $\quad \arg\max_{\theta} \mathcal{L}(\theta|X) = \arg\max_{\theta} \log \mathcal{L}(\theta|X)$ *"Log" likelihood*

Trick #2: finding max (or other extreme values) of a function is simply analyzing the '**stationary points**' of a function. That is, values at which the **derivative** of the function is zero !



MLE Example (contd. 2)

Let's calculate the best fitting $\theta = \{\mu, \sigma^2\}$

$$
\begin{aligned}
\arg\max_{\theta} \mathcal{L}(\theta|X) &= \arg\max_{\theta} \log \mathcal{L}(\theta|X) && \textit{“Log” likelihood} \\
&= \arg\max_{\mu,\sigma^2} \log\left(\prod_{i=1}^{n} p_{\{\mu,\sigma^2\}}(x_i)\right) && \textit{i.i.d.} \\
&= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \log\left(p_{\{\mu,\sigma^2\}}(x_i)\right) \\
&= \arg\max_{\mu,\sigma^2} \sum_{i=1}^{n} \left[\underbrace{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2}}_{g_i(\mu,\sigma^2)} \right]
\end{aligned}
$$

Maximizing μ : $$0 = \nabla_\mu\left(\sum_{i=1}^{n} g_i(\mu,\sigma^2)\right) \implies \mu_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Maximizing σ^2: $$\sigma^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

MLE Example

So, the best fitting **Gaussian model** $p_{\{\mu,\sigma^2\}}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Female height data: 60, 62, 53, 58, … $\in \mathbf{R}$

$$x_1, x_2, \ldots x_n \in \mathcal{X}$$

Is the one with parameters: $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

What about other model classes?

Other popular probability models

Bernoulli model (coin tosses) *Scalar valued*

Multinomial model (dice rolls) *Scalar valued*

Poisson model (rare counting events) *Scalar valued*

Gaussian model (most common phenomenon) *Scalar valued*

Most machine learning data is vector valued!

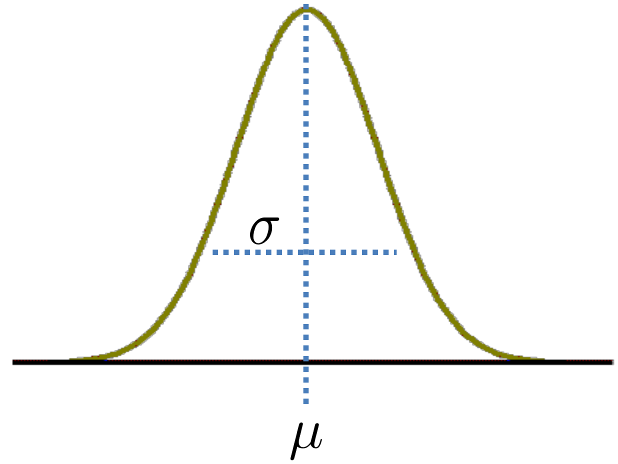
Multivariate Gaussian Model ***Vector valued***

Multivariate version available of other scalar valued models

Multivariate Gaussian

Univariate **R**

$$p_{\{\mu,\sigma^2\}}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$



μ = mean parameter

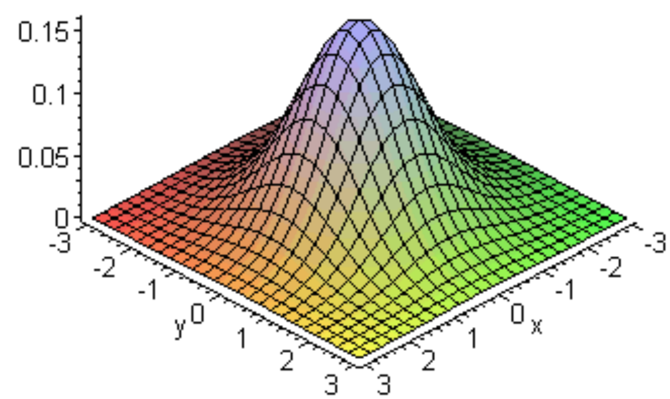
σ^2 = variance parameter > 0

Multivariate **R**^{d}

$$p_{\{\vec{\mu},\Sigma\}}(\vec{x}) := \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^{\mathsf{T}}\Sigma^{-1}(\vec{x}-\vec{\mu}) \right)$$

$\vec{\mu}$ = mean vector

Σ = Covariance matrix (positive definite)



From MLE to Classification

MLE sounds great, how do we use it to do classification using **labelled** data?

$$\hat{f}(\vec{x}) = \arg\max_{y \in \mathcal{Y}} \; P[Y = y | X = \vec{x}]$$

Bayes optimal classifier

$$= \arg\max_{y \in \mathcal{Y}} \; \frac{P[X = \vec{x} | Y = y] \cdot P[Y = y]}{P[X = \vec{x}]}$$

Bayes rule

indep. of y

$$= \arg\max_{y \in \mathcal{Y}} \; P[X = \vec{x} | Y = y] \; \cdot \; P[Y = y]$$

Class conditional probability model

Class Prior

Class prior:

Simply the probability of data sample occurring from a category

Class conditional:

Use a separate probability model individual categories/class-type

We can find the appropriate parameters for the model using MLE!

Classification via MLE Example

Task: learn a classifier to distinguish **`males`** from **`females`**

based on say height and weight measurements

Classifier: $$\hat{f}(\vec{x}) = \underset{y \in \{\text{male}, \text{female}\}}{\arg\max} \; P[X = \vec{x} | Y = y] \cdot P[Y = y]$$

Using **labelled** training data, learn all the parameters:

Learning **class priors**:

$P[Y = \text{male}] =$ fraction of training data labelled as male

$P[Y = \text{female}] =$ fraction of training data labelled as female

Learning **class conditionals**:

$$P[X|Y = \text{male}] = p_{\theta(\text{male})}(X)$$

$$P[X|Y = \text{female}] = p_{\theta(\text{female})}(X)$$

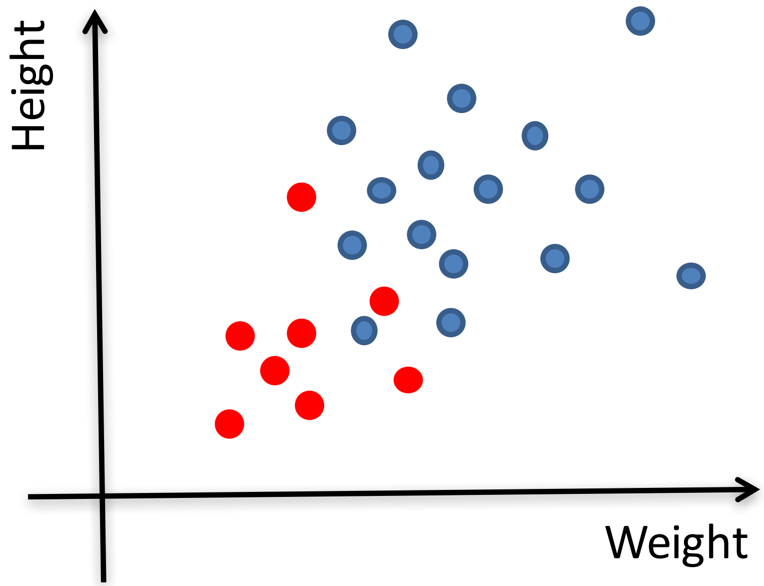
θ(male) = **MLE** using only male data

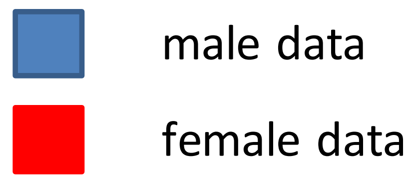
θ(female) = **MLE** using only female data

What are we doing geometrically?

Data geometry:

$$\mathcal{X} = \mathbf{R}^2 = \text{Weight} \times \text{Height}$$

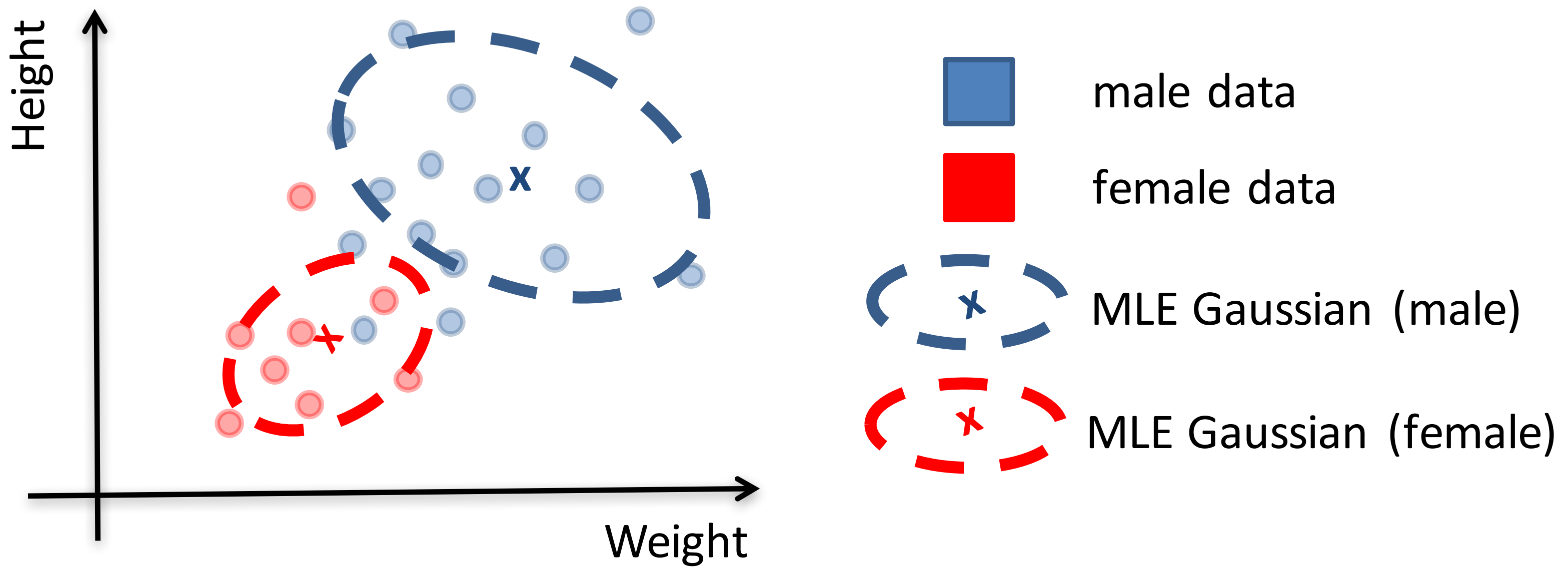


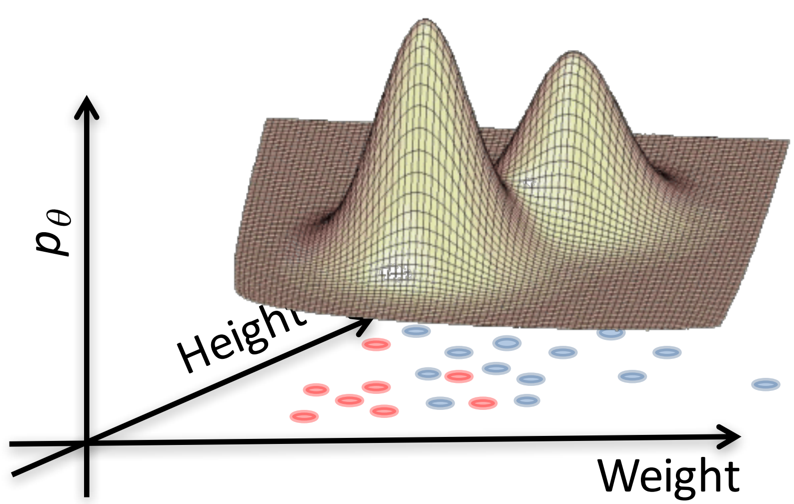


What are we doing geometrically?

Data geometry:

$$\mathcal{X} = \mathbf{R}^2 = \text{Weight} \times \text{Height}$$





Classification via MLE Example

Task: learn a classifier to distinguish **`males`** from **`females`**

based on say height and weight measurements

Classifier: $$\hat{f}(\vec{x}) = \underset{y \in \{\text{male},\text{female}\}}{\arg\max} \; P[X = \vec{x} | Y = y] \cdot P[Y = y]$$

Using **labelled** training data, learn all the parameters:

Learning **class priors**:

$P[Y = \text{male}] =$ fraction of training data labelled as male

$P[Y = \text{female}] =$ fraction of training data labelled as male

Learning **class conditionals**:

$$P[X | Y = \text{male}] = p_{\theta(\text{male})}(X)$$

$$P[X | Y = \text{female}] = p_{\theta(\text{female})}(X)$$

θ(male) = **MLE** using only male data

θ(female) = **MLE** using only female data

Classification via Prob. Models: Variation

Naïve Bayes classifier:

$$\hat{f}(\vec{x}) = \arg\max_{y \in \mathcal{Y}} \; P[X = \vec{x} | Y = y] \; \cdot \; P[Y = y]$$

$$= \arg\max_{y \in \mathcal{Y}} \; \prod_{j=1}^{d} P[X^{(j)} = x^{(j)} | Y = y] \; \cdot \; P[Y = y]$$

$$\vec{x} = \begin{pmatrix} x^{(1)} \\ \vdots \\ x^{(d)} \end{pmatrix}$$

Naïve Bayes assumption: The individual features/measurements are **independent** given the class label

Advantages:

Computationally very simple model. Quick to code.

Disadvantages:

Does not properly capture the interdependence between features, giving bad estimates.

What we learned…

- Maximum Likelihood Estimation
- Learning a classifier via probabilistic modelling
- Naïve Bayes classifier

Questions?