Lecture 15: GMM & Bayesian Learning

Winter 2018

Kai-Wei Chang

CS @ UCLA

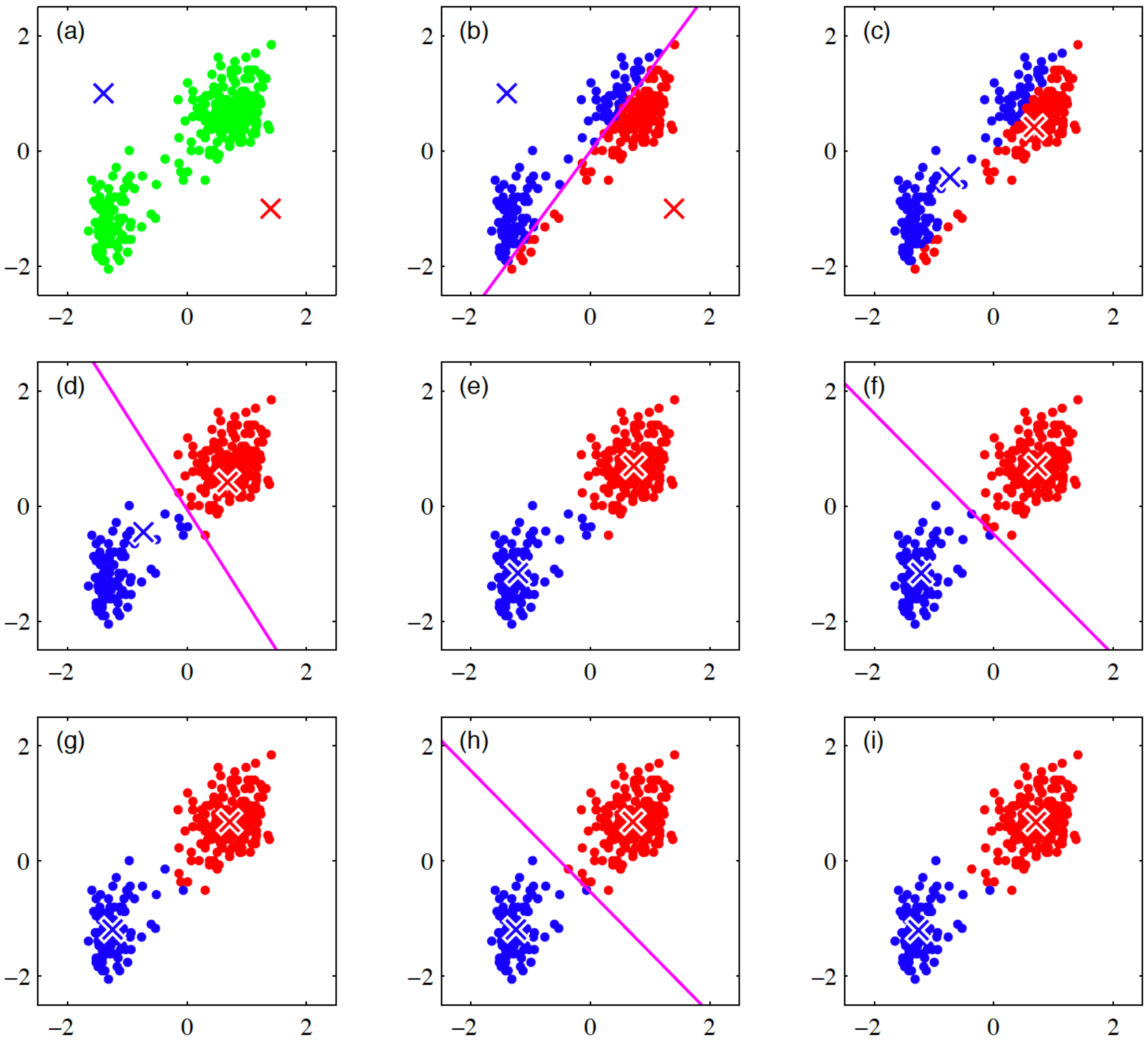
kw+cm146@kwchang.net

The instructor gratefully acknowledges Dan Roth, Vivek Srikuar, Sriram Sankararaman, Fei Sha, Ameet Talwalkar, Eric Eaton, and Jessica Wu whose slides are heavily used, and the many others who made their course material freely available online.

Clustering

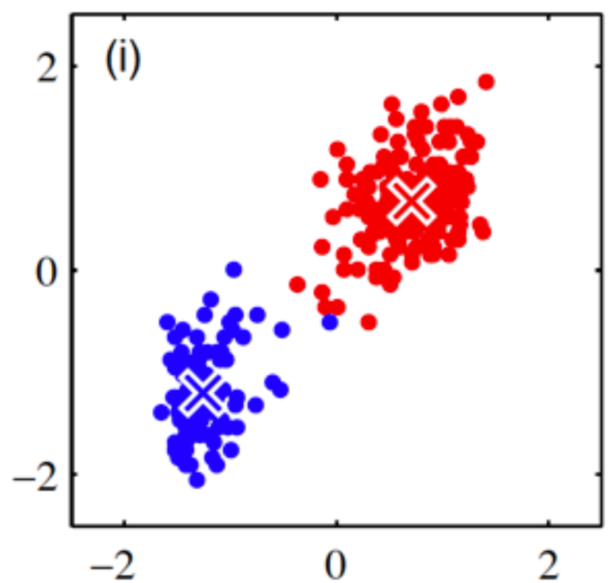
- Cluster students into four groups

Intuition of K-Means



Probabilistic interpretation of clustering?

- Until now, we make a hard assignment for clustering
  - Each point assigns to one cluster
  - Can we allow probability in the assignment?



Recap: Bayes Theorem

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

Short for

$$\forall x, y \quad P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

Bayes Theorem

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

Bayes Theorem

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

Prior probability: What is our belief in Y before we see X?

Bayes Theorem

$$P(Y|X) = \frac{\underbrace{P(X|Y)}_{\text{Likelihood}}\,\underbrace{P(Y)}_{\text{Prior probability}}}{P(X)}$$

Likelihood: What is the likelihood of observing X given a specific Y?

Prior probability: What is our belief in Y before we see X?

Bayes Theorem

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

Posterior probability: What is the probability of Y given that X is observed?

Likelihood: What is the likelihood of observing X given a specific Y?

Prior probability: What is our belief in Y before we see X?

Bayes Theorem

$$\boxed{P(Y|X)} = \frac{\boxed{P(X|Y)}\,\boxed{P(Y)}}{P(X)}$$

Posterior probability: What is the probability of Y given that X is observed?

Likelihood: What is the likelihood of observing X given a specific Y?

Prior probability: What is our belief in Y before we see X?

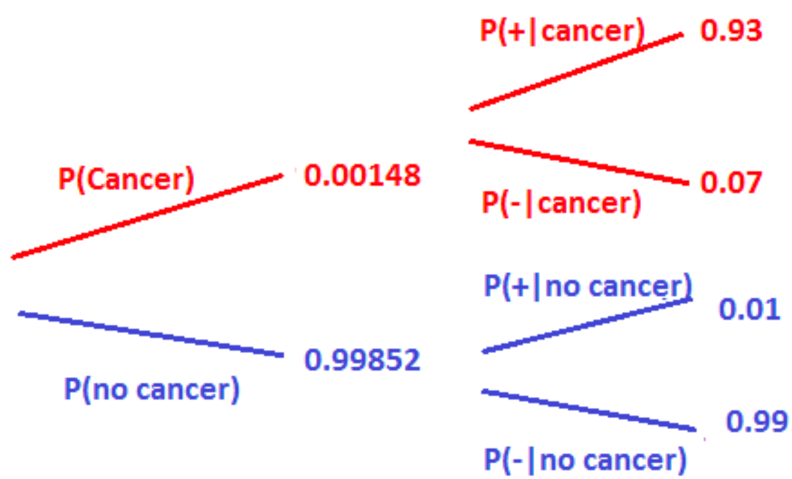
$$\forall x, y \quad P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

$$= \frac{P(X = x|Y = y)P(Y=y)}{\sum_{y'} P(X = x|Y = y')P(Y=y')}$$

Recap: Bayes Theorem Example

- How likely the patient got cancer if the test is positive?

$$\text{P (CANCER} \mid +) = \frac{P(\textit{cancer and}+)}{P(\textit{cancer and}+) + P(\textit{no cancer and}+)} = 0.12$$



Today's lecture

- GMM

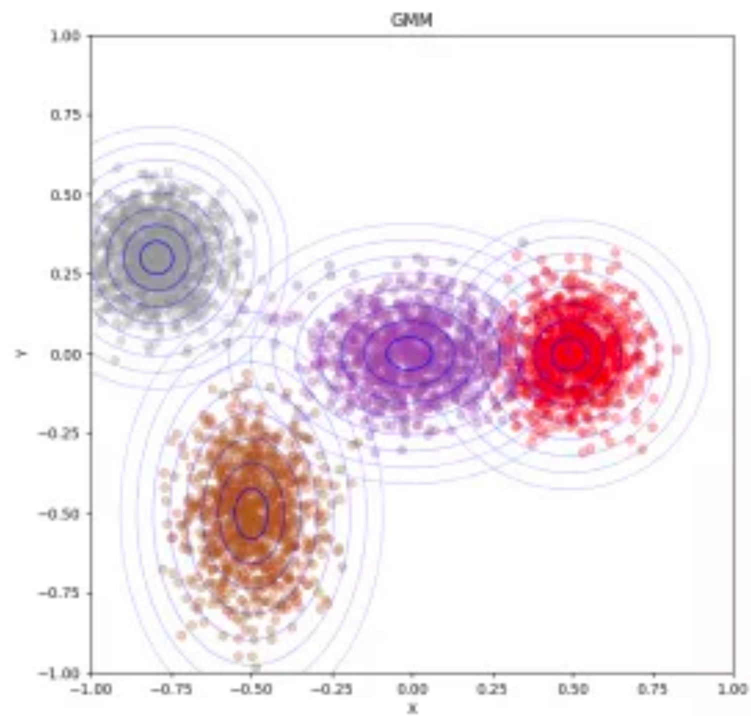
- Bayesian Learning

- Maximum a posteriori and maximum likelihood estimation

- Naïve Bayes

Gaussian mixture models

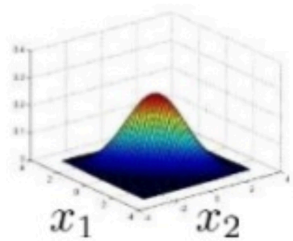
- Intuition: we can model each region with a distinct distribution

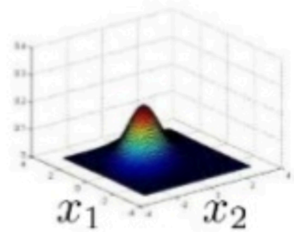


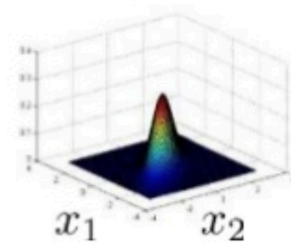
Properties of Gaussian distribution

- Parameters μ, Σ

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$







Parameter fitting:

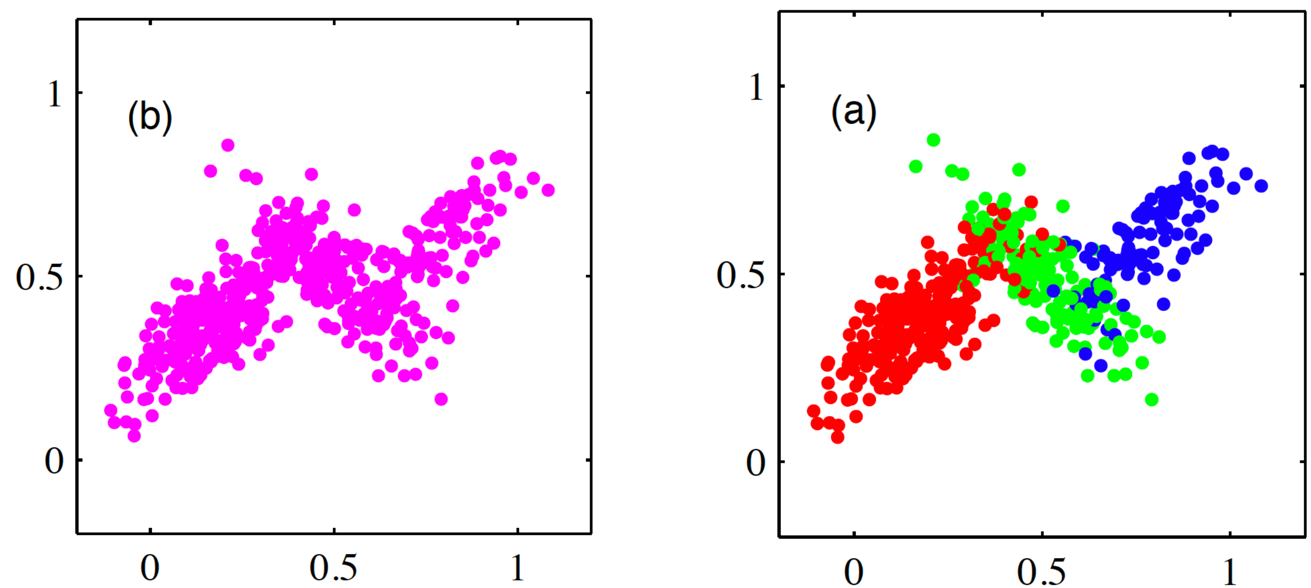
Given training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \qquad \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

Gaussian mixture models

- Assume the probability density function for x as

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \omega_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



Gaussian mixture models: formal definition

A Gaussian mixture model has the following density function for $\boldsymbol{x}$

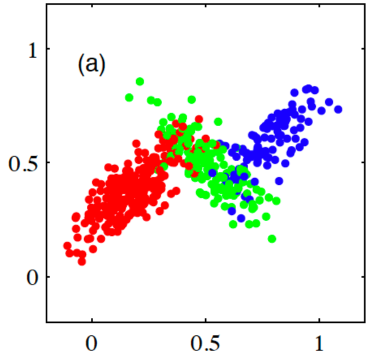
$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \omega_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- K: the number of Gaussians — they are called (mixture) components
- $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$: mean and covariance matrix of the k-th component
- ω_k: mixture weights – they represent how much each component contributes to the final distribution. It satisfies two properties:

$$\forall\ k,\ \omega_k > 0, \quad \text{and} \quad \sum_k \omega_k = 1$$

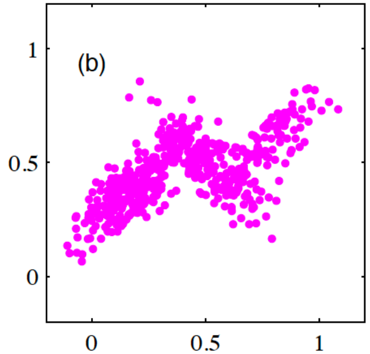
The properties ensure $p(\boldsymbol{x})$ is a properly normalized probability density function.

Example



The conditional distribution between $\boldsymbol{x}$ and z (representing color) are

$$p(\boldsymbol{x}|z = red) = N(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$
$$p(\boldsymbol{x}|z = blue) = N(\boldsymbol{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$
$$p(\boldsymbol{x}|z = green) = N(\boldsymbol{x}|\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$$



The marginal distribution is thus

$$p(\boldsymbol{x}) = p(red)N(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + p(blue)N(\boldsymbol{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + p(green)N(\boldsymbol{x}|\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$$

ω_k

GMM as the marginal distribution P(x) of a joint distribution P(x, z)

Consider the following joint distribution

$$p(\boldsymbol{x}, z) = p(z)p(\boldsymbol{x}|z)$$

where z is a discrete random variable taking values between 1 and K. Denote

$$\omega_k = p(z = k)$$

Now, assume the conditional distributions are Gaussian distributions

$$p(\boldsymbol{x}|z = k) = N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Then, the marginal distribution of $\boldsymbol{x}$ is

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \omega_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Namely, the Gaussian mixture model

Parameter estimation for GMMs

- If cluster assignments are observed $\{z_n\}$ are given
  - We know the cluster of each point
  - Let $\gamma_{nk} = 1$ if instance n belongs to cluster k, otherwise $\gamma_{nk} = 0$
- Then the maximum likelihood estimation is

$$\omega_k = \frac{\sum_n \gamma_{nk}}{\sum_k \sum_n \gamma_{nk}}, \quad \boldsymbol{\mu}_k = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\omega_k = \frac{\sum_n \gamma_{nk}}{\sum_k \sum_n \gamma_{nk}}, \quad \boldsymbol{\mu}_k = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

Since γ_{nk} is binary, the previous solution is nothing but

- For ω_k: count the number of data points whose z_n is k and divide by the total number of data points (note that $\sum_k \sum_n \gamma_{nk} = N$)
- For $\boldsymbol{\mu}_k$: get all the data points whose z_n is k, compute their mean
- For $\boldsymbol{\Sigma}_k$: get all the data points whose z_n is k, compute their covariance matrix

Parameter estimation for GMMs

- When the cluster assignments are not given – the real case
- We use an approach similar to k-means
- Alternative update the cluster assignment γ_{nk} and parameter estimation $\{\omega_k, \mu_k, \Sigma_k\}$

Iterative procedure

- Let θ represent all parameters $\{\omega_k, \mu_k, \Sigma_k\}$

Step 0: initialize $\boldsymbol{\theta}$ with some values (random or otherwise)

Step 1: compute γ_{nk} using the current $\boldsymbol{\theta}$

Step 2: update $\boldsymbol{\theta}$ using the just computed γ_{nk}

Step 3: go back to Step 1

Estimate γ_{nk}

- γ_{nk} the assignment of instance n to cluster k, can be defined as $\gamma_{nk} = P(z_n = k|x_n)$
- Can be computed vis the posterior probability

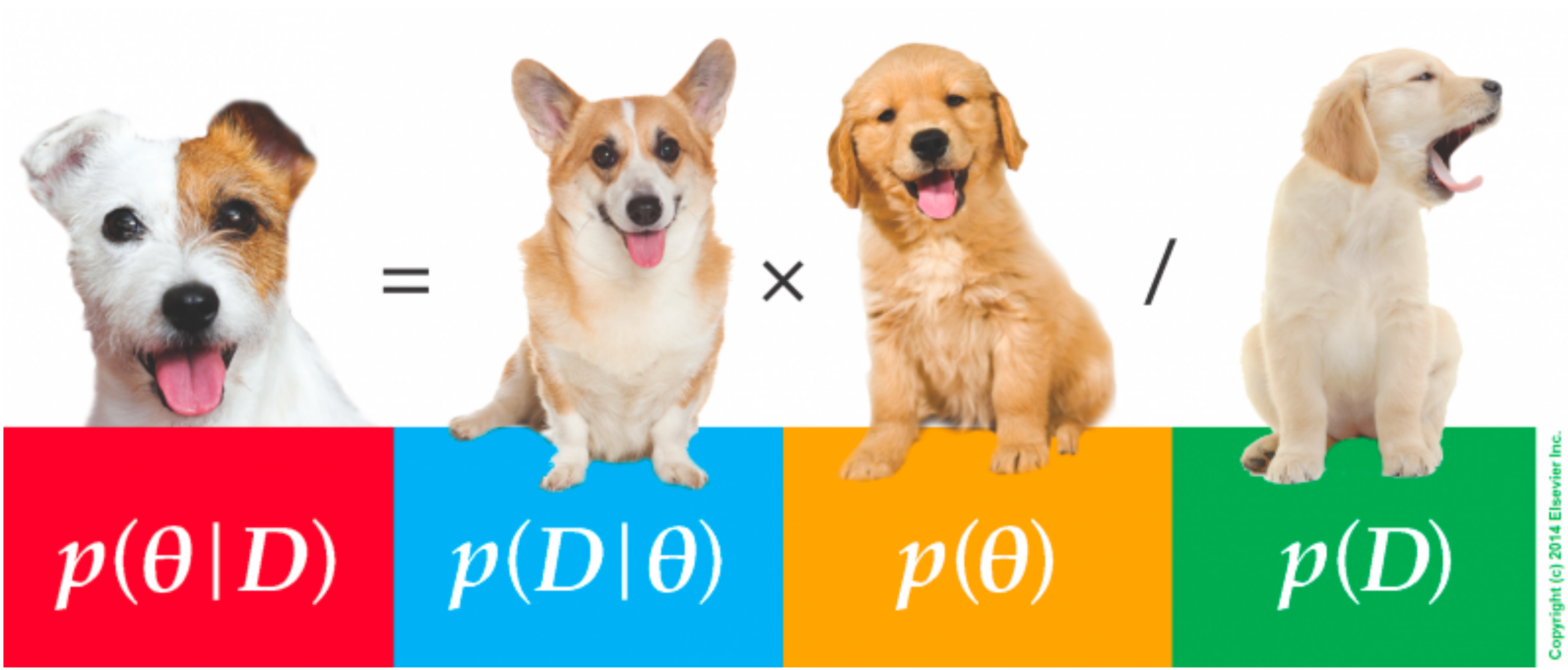
$$p(z_n = k|\boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_n|z_n = k)p(z_n = k)}{p(\boldsymbol{x}_n)} = \frac{\overset{\boxed{N(x|\mu_k, \Sigma_k)}}{p(\boldsymbol{x}_n|z_n = k)}\,\overset{\boxed{\omega_k}}{p(z_n = k)}}{\sum_{k'=1}^{K} p(\boldsymbol{x}_n|z_n = k')p(z_n = k')}$$

Estimate γ_{nk}

- γ_{nk} the assignment of instance n to cluster k, can be defined as $\gamma_{nk} = P(z_n = k|x_n)$
- Can be computed vis the posterior probability

$$p(z_n = k|\boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_n|z_n = k)p(z_n = k)}{p(\boldsymbol{x}_n)} = \frac{\overbrace{p(\boldsymbol{x}_n|z_n = k)}^{\boxed{N(x|\mu_k, \Sigma_k)}}\overbrace{p(z_n = k)}^{\boxed{\omega_k}}}{\sum_{k'=1}^{K} p(\boldsymbol{x}_n|z_n = k')p(z_n = k')}$$

Probabilistic models and Bayesian Learning



Probabilistic Learning

Two different notions of probabilistic learning

- Learning probabilistic concepts
  - The learned concept is a function $c: X \rightarrow [0,1]$
  - $c(x)$ may be interpreted as the probability that the label 1 is assigned to x
  - The learning theory that we have studied before is applicable (with some extensions)

- Bayesian Learning: Use of a probabilistic criterion in selecting a hypothesis
  - The hypothesis can be deterministic, a Boolean function
  - The criterion for selecting the hypothesis is probabilistic

Bayesian Learning: The basics

- Goal: To find the ***best*** hypothesis from some space H of hypotheses, using the observed data D

- Define best = most probable hypothesis in H

- In order to do that, we need to assume a probability distribution over the class H

- We also need to know something about the relation between the data observed and the hypotheses
  - As we will see, we can be Bayesian about other things. e.g., the parameters of the model

Bayesian methods have multiple roles

- Provide practical learning algorithms
- Combining prior knowledge with observed data
  - Guide the model towards something we know
- Provide a conceptual framework
  - For evaluating and analyzing learners

Bayes Theorem

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

$$\Big\downarrow \ \boxed{\text{Short for}}$$

$$\forall x, y \quad P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

Bayes Theorem

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

Posterior probability: What is the probability of Y given that X is observed?

Likelihood: What is the likelihood of observing X given a specific Y?

Prior probability: What is our belief in Y before we see X?

Posterior / Likelihood £ Prior

Bayesian Learning

Given a dataset D, we want to find the <u>best</u> hypothesis h

What does *best* mean?

Bayesian learning uses $P(h\,|D)$, the conditional probability of a hypothesis given the data, to define *best*.

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D)$$

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h

What does *best* mean?

$$P(h|D)$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D)$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Key insight: Both h and D are events.

- D: The event that we observed *this* particular dataset
- h: The event that the hypothesis h is the true hypothesis

<u>So we can apply the Bayes rule here.</u>

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Key insight: Both h and D are events.

- D: The event that we observed *this* particular dataset
- h: The event that the hypothesis h is the true hypothesis

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D) = \frac{P(D|h)\boxed{P(h)}}{P(D)}$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Prior probability of h: Background knowledge. What do we expect the hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Likelihood: What is the probability that this data point (an example or an entire dataset) is observed, given that the hypothesis is h?

Prior probability of h: Background knowledge. What do we expect the hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h

What does *best* mean?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Likelihood: What is the probability that this data point (an example or an entire dataset) is observed, given that the hypothesis is h?

Prior probability of h: Background knowledge. What do we expect the hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.

What is the probability that the data D is observed (independent of any knowledge about the hypothesis)?

Today's lecture

- GMM
- Bayesian Learning
- Maximum a posteriori and maximum likelihood estimation
- Naïve Bayes

Choosing a hypothesis

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$

Choosing a hypothesis

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} P(h|D) \\
&= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\
&= \arg\max_{h \in H} P(D|h)P(h)
\end{aligned}
$$

Choosing a hypothesis

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} P(h|D) \\
&= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\
&= \arg\max_{h \in H} P(D|h)P(h)
\end{aligned}
$$

Posterior / Likelihood £ Prior

Choosing a hypothesis

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

Choosing a hypothesis

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

If we assume that the prior is uniform i.e. $P(h_i) = P(h_j)$, for all h_i, h_j

- Simplify this to get the Maximum Likelihood hypothesis

$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

Often computationally easier to maximize *log likelihood*

Brute force MAP learner

Input: Data D and a hypothesis set H

1. Calculate the posterior probability for each h ∈ H

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

2. Output the hypothesis with the highest posterior probability

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

Maximum Likelihood estimation

Maximum Likelihood estimation (MLE)

$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

What we need in order to define learning:

1. A hypothesis space H
2. A model that says how data D is generated given h

Example: Bernoulli trials

The CEO of a startup hires you for your first consulting job

- *CEO*: My company makes light bulbs. I need to know what is the probability they are faulty.
- *You*: Sure. I can help you out. Are they all identical?
- *CEO*: Yes!
- *You*: Excellent. I know how to help. We need to experiment…

Faulty lightbulbs

The experiment:

Try out 100 lightbulbs

80 work, 20 don’t

You: The probability is P(failure) = 0.2

CEO: But how do you know?

You: Because…

Bernoulli trials

- P(failure) = p, P(success) = 1 – p

- Each trial is i.i.d
  - Independent and identically distributed

Bernoulli trials

- P(failure) = p, P(success) = 1 – p

- Each trial is i.i.d
  - Independent and identically distributed

- You have seen D = {80 work, 20 don't}

$$P(D|p) = \binom{100}{80} p^{80}(1-p)^{20}$$

Bernoulli trials

- P(failure) = p, P(success) = 1 – p

- Each trial is i.i.d
  - Independent and identically distributed

- You have seen D = {80 work, 20 don't}

$$P(D|p) = \binom{100}{80} p^{80}(1-p)^{20}$$

- The most likely value of p for this observation is?

Bernoulli trials

- P(failure) = p, P(success) = 1 – p

- Each trial is i.i.d
  - Independent and identically distributed

- You have seen D = {80 work, 20 don't}

$$P(D|p) = \binom{100}{80} p^{80}(1-p)^{20}$$

- The most likely value of p for this observation is?

$$\underset{p}{\operatorname{argmax}} P(D|p) = \underset{p}{\operatorname{argmax}} \binom{100}{80} p^{80}(1-p)^{20}$$

The "learning" algorithm

Say you have *a* Work and b Not-Work

$$
\begin{aligned}
p_{best} &= \underset{p}{\operatorname{argmax}}\, P(D|h) \\
&= \underset{p}{\operatorname{argmax}} \log P(D\,|h) \\
&= \underset{p}{\operatorname{argmax}} \log\left(\binom{a+b}{a} p^a (1-p)^b\right) \\
&= \underset{p}{\operatorname{argmax}}\, a \log p + b \log(1-p)
\end{aligned}
$$

Calculus 101: Set the derivative to zero

$$P_{best} = b/(a + b)$$

The "learning" algorithm

Say you have *a* Work and b Not-Work

$$
\begin{aligned}
p_{best} &= \underset{p}{\operatorname{argmax}} P(D|h) \\
&= \underset{p}{\operatorname{argmax}} \log P(D\,|h) \\
&= \underset{p}{\operatorname{argmax}} \log \left(\binom{a+b}{a} p^a (1-p)^b \right) \\
&= \underset{p}{\operatorname{argmax}}\, a \log p + b \log(1-p)
\end{aligned}
$$

Log likelihood

Calculus 101: Set the derivative to zero

$$P_{best} = b/(a + b)$$

The “learning” algorithm

Say you have *a* Work and b Not-Work

$$
\begin{aligned}
p_{best} &= \underset{p}{\operatorname{argmax}} P(D|h) \\
&= \underset{p}{\operatorname{argmax}} \log P(D\,|h) \\
&= \underset{p}{\operatorname{argmax}} \log\left(\binom{a+b}{a} p^a (1-p)^b\right) \\
&= \underset{p}{\operatorname{argmax}}\, a \log p + b \log(1-p)
\end{aligned}
$$

Log likelihood

Calculus 101: Set the derivative to zero

$$P_{best} = b/(a + b)$$

MAP estimation

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$h_{MAP} = \arg\max_{h \in H} P(D|h) \boxed{P(h)}$$

If we assume that the prior is uniform i.e. $P(h_i) = P(h_j)$, for all h_i, h_j

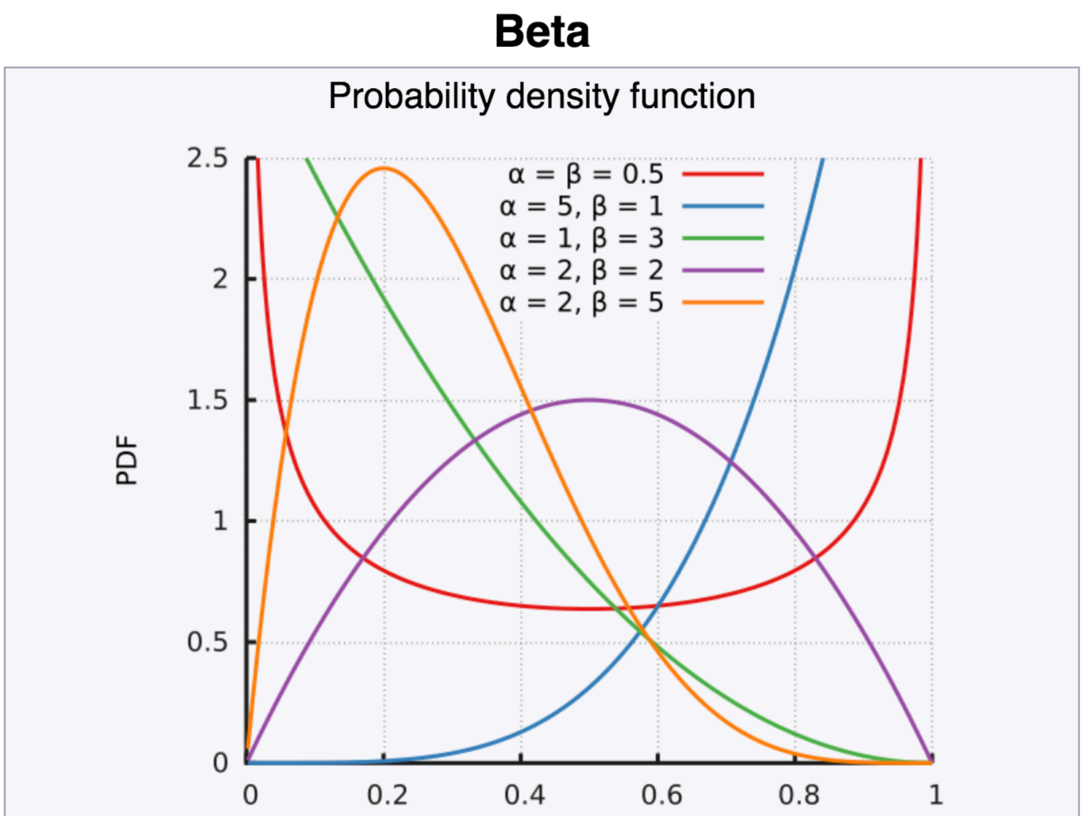
- Simplify this to get the Maximum Likelihood hypothesis

$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

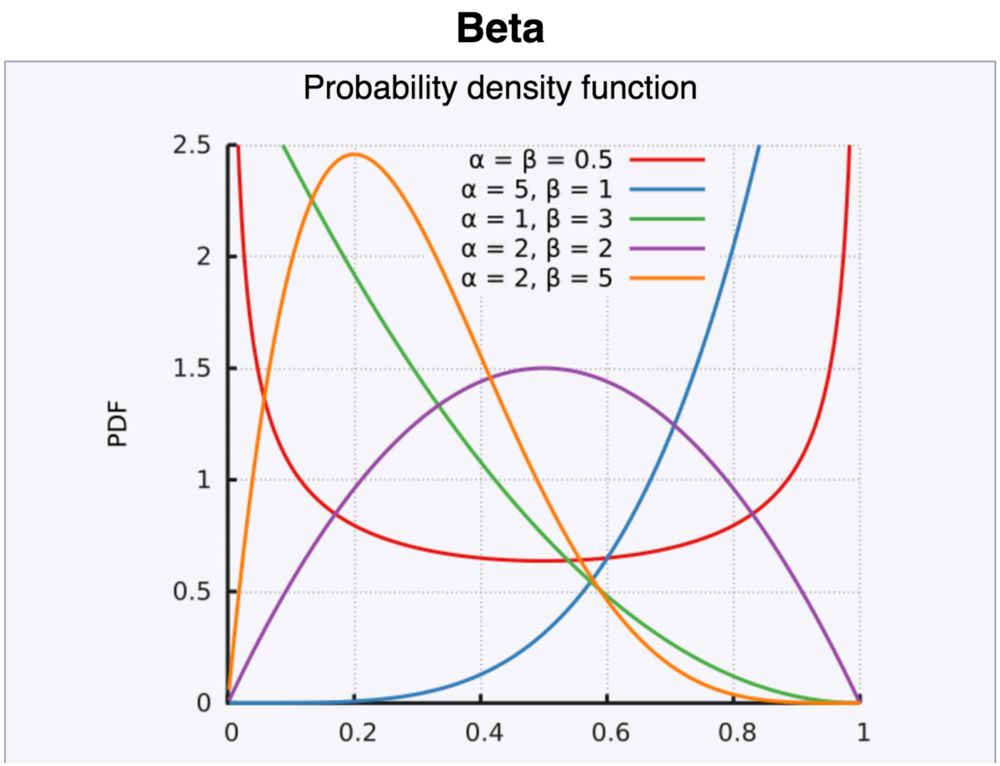
Often computationally easier to maximize *log likelihood*

Prior distribution

- The boss has a prior belief of the ditribution of faulty lightbulb



Beta distribution



$$
\begin{aligned}
f(x;\alpha,\beta) &= \mathrm{constant}\cdot x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}
\end{aligned}
$$

MAP for Bernoulli trials

p is the parameter for hypothesis

- P(failure) = p, P(success) = 1 – p

- Each trial is i.i.d
  - Independent and identically distributed

- You have seen D = {80 work, 20 don't}

$$P(D|p) = \binom{100}{80} p^{80}(1-p)^{20}$$

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

MAP estimation

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

$$P(D|p) = \binom{100}{80} p^{80}(1-p)^{20}$$

$$P(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$p_{best} = \underset{p}{\operatorname{argmax}}\, P(D|h)\, P(h)$$

$$= \underset{p}{\operatorname{argmax}} \log P(D\,|h) + \log P(h)$$

$$= \underset{p}{\operatorname{argmax}} \log\left(\frac{\binom{a+b}{a}}{B(\alpha,\beta)} p^{a}(1-p)^{b} p^{\alpha-1}(1-p)^{\beta-1}\right)$$

$$= \underset{p}{\operatorname{argmax}}\ (a+\alpha-1)\log p + (b+\beta-1)\log(1-p)$$

MAP estimation

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

$$P(D|p) = \binom{100}{80} p^{80}(1-p)^{20}$$

$$P(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$p_{best} = \underset{p}{\text{argmax}}\, P(D|h)\, P(h)$$

$$= \underset{p}{\text{argmax}} \log P(D\,|h) + \log P(h)$$

$$= \underset{p}{\text{argmax}} \log \left(\frac{\binom{a+b}{a}}{B(\alpha,\beta)} p^{a}(1-p)^{b} p^{\alpha-1}(1-p)^{\beta-1} \right)$$

$$= \underset{p}{\text{argmax}}\ (a+\alpha-1)\log p + (b+\beta-1)\log(1-p)$$

MAP v.s. MLE

- MLE:

$$\underset{p}{\operatorname{argmax}}\, a\log p + b\log(1-p)$$

$$\Rightarrow p_{best} = \frac{a}{a+b}$$

- MAP

$$\underset{p}{\operatorname{argmax}}\ (a+\alpha-1)\log p + (b+\beta-1)\log(1-p)$$

$$\Rightarrow p_{best} = \frac{a+\alpha-1}{a+b+\alpha+\beta-2}$$

MAP v.s. MLE

- MAP

  $$\underset{p}{\operatorname{argmax}}\ (a+\alpha-1)\log p + (b+\beta-1)\log(1-p)$$

  $$\Rightarrow p_{best} = \frac{a+\alpha-1}{a+b+\alpha+\beta-2}$$

- Let $\alpha = 100,\ \ \beta = 10$
- $a = 10, b = 20 \Rightarrow p_{best} \approx 0.79$
- $a = 1000, b = 2000 \Rightarrow p_{best} \approx 0.36$
- $a = 100{,}000, b = 200{,}000 \Rightarrow p_{best} \approx 0.33$

Advanced topic (not cover in exam) MAP for logistic regression

Let's get back to the MLE for logistic regression

- Training data
  - S = $\{(\mathbf{x}_i, \mathbf{y}_i)\}$, m examples
- What we want
  - Find a **w** such that P(S | **w**) is maximized
  - We know that our examples are drawn independently and are identically distributed (i.i.d)
  - How do we proceed?

Maximum likelihood estimation

$$\underset{\mathbf{w}}{\operatorname{argmax}}\, P(S|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

The usual trick: Convert products to sums by taking log

Recall that this works only because log is an increasing function and the maximizer will not change

Maximum likelihood estimation

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\Bigg\downarrow \boxed{\text{Equivalent to solving}}$$

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Maximum likelihood estimation

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

But (by definition) we know that

$$P(y|\mathbf{w}, \mathbf{x}) = \sigma(y_i \mathbf{w}^T \mathbf{x}_i) = \frac{1}{1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)}$$

$$P(y|\mathbf{w},\mathbf{x}) = \frac{1}{1+\exp(-\mathrm{y_i}\mathbf{w}^T\mathbf{x}_i)}$$

Maximum likelihood estimation

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i,\mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i,\mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i))$$

$$P(y|\mathbf{w},\mathbf{x}) = \frac{1}{1+\exp(-\mathrm{y_i}\mathbf{w}^T\mathbf{x}_i)}$$

Maximum likelihood estimation

$$\boxed{\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w})} = \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i,\mathbf{w})$$

The goal: Maximum likelihood training of a discriminative probabilistic classifier under the logistic model for the posterior distribution.

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i,\mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i))$$

$$P(y|\mathbf{w},\mathbf{x}) = \frac{1}{1+\exp(-\mathrm{y_i}\mathbf{w}^T\mathbf{x}_i)}$$

Maximum likelihood estimation

$$\underset{\mathbf{w}}{\operatorname{argmax}}\, P(S|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i,\mathbf{w})$$

The goal: Maximum likelihood training of a discriminative probabilistic classifier under the logistic model for the posterior distribution.

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i,\mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i))$$

Equivalent to: Training a linear classifier by minimizing the ***logistic loss***.

Maximum a posteriori estimation

We could also add a prior on the weights

Suppose each weight in the weight vector is drawn independently from the normal distribution with zero mean and standard deviation σ

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

MAP estimation for logistic regression

Maximum likelihood estimation

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i=1}^{m} -\log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right)$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Let us work through this procedure again to see what changes

MAP estimation for logistic regression

Maximum likelihood estimation

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i=1}^{m} -\log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right)$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Let us work through this procedure again to see what changes

What is the goal of MAP estimation? (In maximum likelihood, we maximized the likelihood of the data)

MAP estimation for logistic regression

Maximum likelihood estimation

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i=1}^{m} -\log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right)$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

What is the goal of MAP estimation? (In maximum likelihood, we maximized the likelihood of the data)

To maximize the posterior probability of the model given the data (i.e. to find the most probable model, given the data)

$$P(\mathbf{w}|S) \propto P(S|\mathbf{w})P(\mathbf{w})$$

MAP estimation for logistic regression

Maximum likelihood estimation

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i=1}^{m} -\log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right)$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\operatorname*{argmax}_{\mathbf{w}} P(\mathbf{w}|S) = \operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) P(\mathbf{w})$$

MAP estimation for logistic regression

Maximum likelihood estimation

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i=1}^{m} -\log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right)$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) P(\mathbf{w})$$

Take log to simplify

$$\max_{\mathbf{w}} \log P(S|\mathbf{w}) + \log P(\mathbf{w})$$

MAP estimation for logistic regression

Maximum likelihood estimation

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i=1}^{m} -\log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right)$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) P(\mathbf{w})$$

Take log to simplify

$$\max_{\mathbf{w}} \log P(S|\mathbf{w}) + \log P(\mathbf{w})$$

We have already expanded out the first term.

$$\sum_{i}^{m} -\log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))$$

MAP estimation for logistic regression

Maximum likelihood estimation

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i=1}^{m} -\log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right)$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) P(\mathbf{w})$$

Take log to simplify

$$\max_{\mathbf{w}} \log P(S|\mathbf{w}) + \log P(\mathbf{w})$$

Expand the log prior

$$\sum_{i}^{m} -\log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i) + \sum_{j=1}^{d} \frac{-w_i^2}{\sigma^2} + constants$$

MAP estimation for logistic regression

Maximum likelihood estimation

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i=1}^{m} -\log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right)$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) P(\mathbf{w})$$

Take log to simplify

$$\max_{\mathbf{w}} \log P(S|\mathbf{w}) + \log P(\mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)) + \sum_{j=1}^{d} \frac{-w_i^2}{\sigma^2} + constants$$

MAP estimation for logistic regression

Maximum likelihood estimation

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i=1}^{m} -\log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right)$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\underset{\mathbf{w}}{\operatorname{argmax}}\, P(S|\mathbf{w})P(\mathbf{w})$$

Take log to simplify

$$\max_{\mathbf{w}} \log P(S|\mathbf{w}) + \log P(\mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i) - \frac{1}{\sigma^2} \mathbf{w}^T \mathbf{w}$$

MAP estimation for logistic regression

Maximum likelihood estimation

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

$\boxed{\text{Equivalent to solving}}$

$$\max_{\mathbf{w}} \sum_{i=1}^{m} -\log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right)$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w})P(\mathbf{w})$$

Take log to simplify

$$\max_{\mathbf{w}} \log P(S|\mathbf{w}) + \log P(\mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i) - \frac{1}{\sigma^2} \mathbf{w}^T \mathbf{w}$$

Maximizing a negative function is the same as minimizing the function

Learning a logistic regression classifier

Learning a logistic regression classifier is equivalent to solving

$$\min_{\mathbf{w}} \sum_{i}^{m} \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i) + \frac{1}{\sigma^2} \mathbf{w}^T \mathbf{w}$$

Today's lecture

- GMM
- Bayesian Learning
- Maximum a posteriori and maximum likelihood estimation
- Naïve Bayes

Where are we?

We have seen Bayesian learning

- Using a probabilistic criterion to select a hypothesis
- Maximum a posteriori and maximum likelihood learning
  - Question: What is the difference between them?

We could also learn functions that predict probabilities of outcomes

- Different from using a probabilistic criterion to learn

Maximum a posteriori (MAP) prediction as opposed to MAP learning

MAP prediction

Let's be use the Bayes rule for predicting y given an input $\mathbf{x}$

$$P(Y = y|X = \mathbf{x}) = \frac{P(X = \mathbf{x}|Y = y)P(Y = y)}{P(X = \mathbf{x})}$$

Posterior probability of label being y for this input x

MAP prediction

Let's be use the Bayes rule for predicting y given an input $\mathbf{x}$

$$P(Y=y|X=\mathbf{x}) = \frac{P(X=\mathbf{x}|Y=y)P(Y=y)}{P(X=\mathbf{x})}$$

Predict y for the input $\mathbf{x}$ using

$$\arg\max_y \frac{P(X=\mathbf{x}|Y=y)P(Y=y)}{P(X=\mathbf{x})}$$

MAP prediction

Let's be use the Bayes rule for predicting y given an input $\mathbf{x}$

$$P(Y = y|X = \mathbf{x}) = \frac{P(X = \mathbf{x}|Y = y)P(Y = y)}{P(X = \mathbf{x})}$$

Predict y for the input $\mathbf{x}$ using

$$\arg\max_{y} P(X = \mathbf{x}|Y = y)P(Y = y)$$

MAP prediction

Don't confuse with *MAP learning*: finds hypothesis by

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

Let's be use the Bayes rule for predicting y given an input $\mathbf{x}$

$$P(Y = y|X = \mathbf{x}) = \frac{P(X = \mathbf{x}|Y = y)P(Y = y)}{P(X = \mathbf{x})}$$

Predict y for the input $\mathbf{x}$ using

$$\arg\max_{y} P(X = \mathbf{x}|Y = y)P(Y = y)$$

MAP prediction

Predict y for the input $\mathbf{x}$ using

$$\arg\max_y P(X = \mathbf{x} | Y = y) P(Y = y)$$

Likelihood of observing this input $\mathbf{x}$ when the label is y

Prior probability of the label being y

All we need are these two sets of probabilities

Example: Tennis

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Without any other information, what is the prior probability that I should play tennis?

Example: Tennis

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Without any other information, what is the prior probability that I should play tennis?

Likelihood

Temperature	Wind	P(T, W |Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

On days that I do play tennis, what is the probability that the temperature is T and the wind is W?

Temperature	Wind	P(T, W |Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

On days that I don't play tennis, what is the probability that the temperature is T and the wind is W?

Example: Tennis again

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Likelihood

Temperature	Wind	P(T, W |Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

Temperature	Wind	P(T, W |Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

Input:
Temperature = Hot (H)
Wind = Weak (W)

Should I play tennis?

Example: Tennis again

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Likelihood

Temperature	Wind	P(T, W |Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

Temperature	Wind	P(T, W |Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

Input:
Temperature = Hot (H)
Wind = Weak (W)

Should I play tennis?

argmax_y P(H, W | play?) P (play?)

Example: Tennis again

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Likelihood

Temperature	Wind	P(T, W |Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

Temperature	Wind	P(T, W |Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

Input:
Temperature = Hot (H)
Wind = Weak (W)

Should I play tennis?

argmax_y P(H, W | play?) P (play?)

P(H, W | Yes) P(Yes) = 0.4 £ 0.3
= 0.12

P(H, W | No) P(No) = 0.1 £ 0.7
= 0.07

Example: Tennis again

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Likelihood

Temperature	Wind	P(T, W |Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

Temperature	Wind	P(T, W |Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

Input:
Temperature = Hot (H)
Wind = Weak (W)

Should I play tennis?

argmax_y P(H, W | play?) P (play?)

P(H, W | Yes) P(Yes) = 0.4 £ 0.3
= 0.12

P(H, W | No) P(No) = 0.1 £ 0.7
= 0.07

MAP prediction = Yes