Additional models to consider for the final project

Naïve Bayes classifier

These slides introduce the assumptions behind and formulation of the Naïve Bayes classifier

- For additional reading please see APM section 13.6

- The slides also mention Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)
- For more information on LDA and QDA please see:
  - ISL section 4.4
  - APM section 12.3 (LDA) and section 13.1 (QDA)

Let’s reconsider the classification task

- Rather than viewing the problem as the class based or conditioned on the inputs:

$$p(y_n = 1 \mid \mathbf{x}_n)$$

- Let’s view the problem as the inputs conditioned on the class:

$$p(\mathbf{x}_n \mid y_n = 1)$$

- Or if there are more than two classes, conditioned on class l:

$$p(\mathbf{x}_n \mid y_n = l)$$

This setup is known as the <u>**generative**</u> approach

- Remember that there are, in general, D inputs.
- $p(\mathbf{x}_n \mid y_n = l)$ is therefore a joint density between potentially many variables!

To write out the joint density we need to consider two things

- Are the variables (the inputs) related?

- The specific distribution associated with the variables.

To write out the joint density we need to consider two things

- Are the variables (the inputs) related?

- The specific distribution associated with the variables.

- LDA and QDA handle both aspects by assuming the inputs can be modeled with a MVN!

Another approach is to assume all inputs are <u>conditionally independent</u> given the class

- The joint density can therefore be factored into the product of D densities:

$$\boxed{p(\mathbf{x}_n \mid y_n = l, \boldsymbol{\theta}) = \prod_{d=1}^{D} \left(p\left(x_{n,d} \mid y_n = l, \boldsymbol{\theta}_{d,l}\right) \right)}$$

Another approach is to assume all inputs are <u>conditionally independent</u> given the class

- The joint density can therefore be factored into the product of D densities:

$$\boxed{p(\mathbf{x}_n \mid y_n = l, \boldsymbol{\theta}) = \prod_{d=1}^{D} \Big(p\big(x_{n,d} \mid y_n = l, \boldsymbol{\theta}_{d,l}\big) \Big)}$$

Model formulation is known as NAÏVE BAYES

The term “naïve” is used to represent that we do not think the assumption is true

- We do not actually believe the inputs are all conditionally independent given the class.
- It is a very useful simplification!
- Surprisingly enough, this assumption yields reasonable models!

Besides simplifying the math, the assumption allows combining different distributions!

- We do not need to apply the same distribution type to all inputs.

- For example, continuous inputs can use Gaussians, binary inputs can use Bernoulli distributions, categorical inputs with more than two classes can use multinoulli (categorical) distributions.

- Can even use non-parametric density approaches for complex distributions.

- Naïve Bayes is quite flexible!

The model parameters, $\boldsymbol{\theta}$, are dictated by the selected distribution types.

For continuous inputs with Gaussian distributions:

$$p\left(x_{n,d} \mid y_n = l, \boldsymbol{\theta}_{d,l}\right) = \text{normal}\left(x_{n,d} \mid \mu_{d,l}, \sigma_{d,l}\right)$$

Each continuous input therefore has 2 parameters per class: $\mu_{d,l}$ & $\sigma_{d,l}$.

For binary inputs with Bernoulli distributions:

$$p\left(x_{n,d} \mid y_n = l, \theta_{d,l}\right) = \text{Bernoulli}\left(x_{n,d} \mid \mu_{d,l}\right)$$

Each binary input has 1 parameter, $\mu_{d,l}$.

n-th observation's likelihood: consider the joint density between the inputs and response

$$p(\mathbf{x}_n, y_n \mid \boldsymbol{\theta}) = p(y_n \mid \boldsymbol{\mu}) \prod_{d=1}^{D} \left(p\left(x_{n,d} \mid \boldsymbol{\theta}_d\right) \right)$$

- Rewrite using:
- Indicator or dummy variables to represent the class:
  - $y_{n,l} = 1$, if the n-th observation is the l-th class, $y_{n,l} = 0$ otherwise
- The probability of each class is denoted μ_l

$\boldsymbol{n}$-th observation's likelihood: consider the joint density between the inputs and response

$$p(\mathbf{x}_n, y_n \mid \boldsymbol{\theta}) = \prod_{l=1}^{L} \left(\mu_l^{y_{n,l}}\right) \prod_{d=1}^{D} \left(\prod_{l=1}^{L} \left(p\left(x_{n,d} \mid \boldsymbol{\theta}_{d,l}\right)^{y_{n,l}} \right) \right)$$

- Total number of unknowns:
- Unknown class probabilities $\boldsymbol{\mu} = \{\mu_{l=1}, \ldots, \mu_l, \ldots, \mu_L\}$
- Input parameters per class $\boldsymbol{\theta}_{d,l}$

The complete log-likelihood requires summing over all N observations

- Remember, probability is essentially counting…so let's do some counting!

- The number of times the l-th class is observed: N_l

The complete log-likelihood is:

$$\log[p(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\theta})] = \sum_{l=1}^{L} (N_l \log[\mu_l]) + \sum_{d=1}^{D} \left(\sum_{l=1}^{L} \left(\sum_{y_{n,l}=1} (\log[p(x_{n,d} \mid \boldsymbol{\theta}_{d,l})]) \right) \right)$$

If there are N total observations…

- What do you think the Maximum Likelihood Estimate (MLE) is for each class probability, $\hat{\mu}_l$?

If there are N total observations…

- What do you think the Maximum Likelihood Estimate (MLE) is for each class probability, $\hat{\mu}_l$?

$$\boxed{\hat{\mu}_l = \frac{N_l}{N}}$$

The input-class parameters depend on the distribution associated with each input

- If all inputs are Binary variables, all input-class distributions are Bernoulli distributions:

$$p\left(x_{n,d} \mid y_{n,l} = 1, \theta_{d,l}\right) = \text{Bernoulli}\left(x_{n,d} \mid \mu_{d,l}\right)$$

- How can we calculate the MLE on each $\mu_{d,l}$ parameter?

- Number of times the l-th class was observed: N_l
- Number of times the d-th input was observed associated with the l-th class: $N_{d,l}$

The MLE on each $\mu_{d,l}$ is just more counting!

$$\boxed{\hat{\mu}_{d,l} = \frac{N_{d,l}}{N_l}}$$

Predictions with a Naïve Bayes classifier based on parameter MLEs

- We observe a new input, $\mathbf{x}_*$, what's the probability of the l-th class?

- Continue assuming all inputs are binary variables.

$$p(y_* = l \mid x_*, X, y, \hat{\theta}) \propto \hat{\mu}_l \prod_{d=1}^{D} \left(\mathrm{Bernoulli}(x_{*,d} \mid \hat{\mu}_{d,l})\right)$$

- The predicted class is then the class with the highest predicted probability.

Is Naïve Bayes…Bayesian?

- The model is based on conditional probability rules.

- But what makes a model…Bayesian?

Is Naïve Bayes…Bayesian?

- The model is based on conditional probability rules.
- But what makes a model…Bayesian?
- We did not assign PRIOR distributions to the parameters!

For the case of all binary input variables

- We can assign Beta distributions as the priors on all $\mu_{d,l}$ parameters.

- We can assign Dirichlet priors on all class probabilities, μ_l.

- The factored likelihood is a Multinomial likelihood for the class and a Bernoulli likelihood for the input…what are the posterior distributions?

For the case of all binary input variables

- We can assign Beta distributions as the priors on all $\mu_{d,l}$ parameters.
- We can assign Dirichlet priors on all class probabilities, μ_l.
- The factored likelihood is a Multinomial likelihood for the class and a Bernoulli likelihood for the input…what are the posterior distributions?

Dirichlet and Betas!!