EE5907/EE5027 Week 3: Univariate Gaussian + Naive Bayes

BT Thomas Yeo

ECE, CSC, CIRC, N.1, HMS

Last Week Recap

- Probabilistic Signal Detection / Estimation
  - MAP & ML estimation
  - Two approaches of modeling posterior distribution: Generative & Discriminative Models
  - Parameters of probabilistic models can themselves be treated as random variables and estimated accordingly
- Conjugate priors
  - Beta-Binomial generative model (binomial parameters are beta distributed)
  - Dirichlet-Multinomial generative models (multinomial parameters are dirichlet distributed)
- Predicting future coin toss / dice rolls
  - Posterior predictive distribution, Plug-In Approximation, Black Swan

This Week

- Last week, binomial and multinomial are discrete random variables,
  - This week, look at continuous example: univariate Gaussian
- Then we will then combine what we have learned last lecture and this lecture to study our first generative classifier: naives Bayes

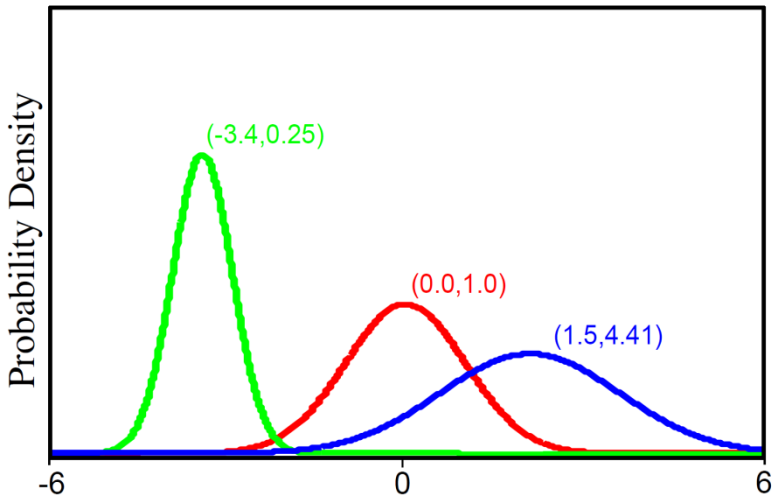
Univariate Gaussian

Univariate Gaussian Distribution

$$Pr(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-0.5(x-\mu)^2/\sigma^2\right]$$

For short we write:

$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$



Univariate normal distribution describes single continuous variable.

Takes 2 parameters μ and $\sigma^2>0$

ML estimation of μ, σ^2 given training data D = $\{x_1,\ldots,x_N\}$

$$(\hat{\mu}, \hat{\sigma}^2) \triangleq \underset{\mu,\sigma^2}{\operatorname{argmax}}\, p(x_1, \cdots, x_N|\mu, \sigma^2)$$

conditional independence

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

Because log is monotonic

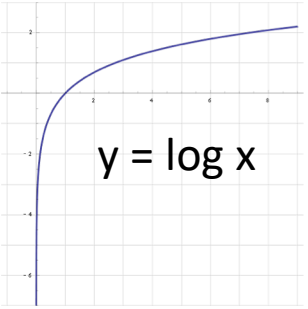
$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \log \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

$$\log(ab) = \log a + \log b$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(x_n|\mu, \sigma^2)$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{n=1}^{N} \left[-\frac{(x_n-\mu)^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2} \right]$$

$$p(x_n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$



ML estimation of μ, σ^2 given training data D = $\{x_1,\ldots,x_N\}$

- From previous slide: $(\hat{\mu}, \hat{\sigma}^2) = \text{argmax}_{\mu,\sigma^2} \sum_{n=1}^{N} \left[-\frac{(x_n-\mu)^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2} \right]$

- Differentiate w.r.t. μ and set to 0:

$$\frac{\partial L}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\sum_{n=1}^{N} -\frac{(x_n-\mu)^2}{2\sigma^2} \right) = \sum_{n=1}^{N} \frac{(x_n-\mu)}{\sigma^2} = 0$$

$$\implies \hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Differentiate w.r.t. σ and set to 0:

$$\frac{\partial L}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left(\sum_{n=1}^{N} -\frac{(x_n-\mu)^2}{2\sigma^2} - N\log\sigma \right) = \sum_{n} \frac{(x_n-\mu)^2}{\sigma^3} - \frac{N}{\sigma} = 0$$

$$\implies \hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\hat{\mu})^2$$

Modeling μ, σ^2 as random variables using Normal Inverse Gamma Distribution

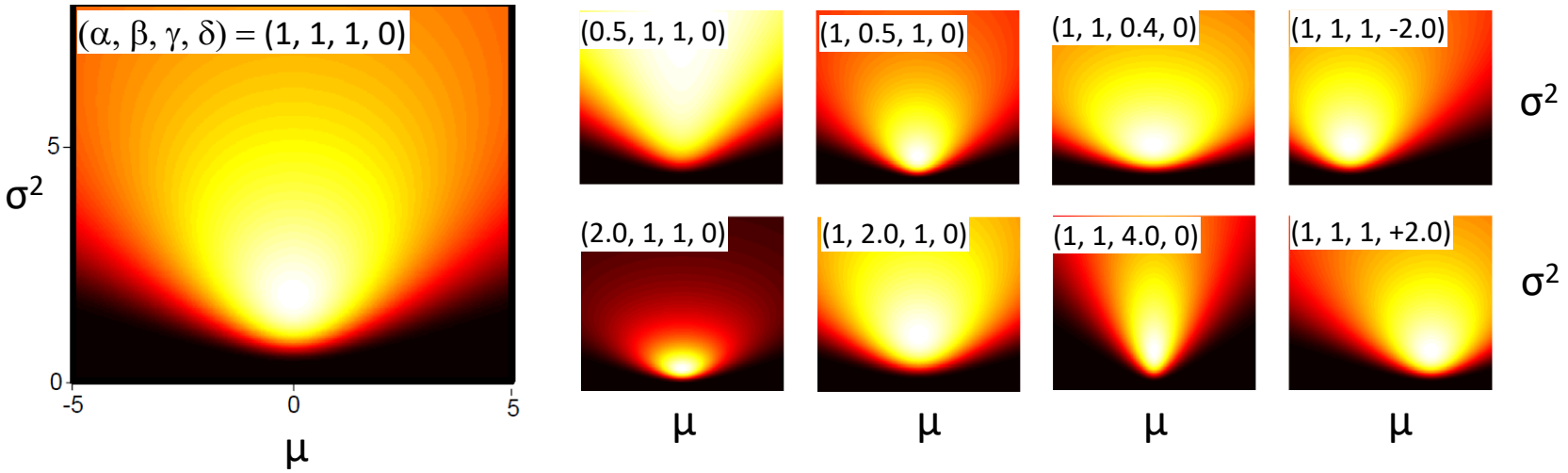
Defined on 2 variables μ and $\sigma^2>0$

$$p(\mu,\sigma^2)=\frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}}\frac{\beta^\alpha}{\Gamma[\alpha]}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left[-\frac{2\beta+\gamma(\delta-\mu)^2}{2\sigma^2}\right]$$

or for short

$$p(\mu,\sigma^2)=\text{NormInvGam}(\mu,\sigma^2|\alpha,\beta,\gamma,\delta)$$

Four parameters $\alpha>0$, $\beta>0$, $\gamma>0$ and δ.



MAP estimation of μ, σ^2 given training data $D = \{x_1,\ldots,x_N\}$

$$(\hat{\mu},\hat{\sigma}^2) \stackrel{\triangle}{=} \operatorname*{argmax}_{\mu,\sigma^2} p(\mu,\sigma^2|x_1,\cdots,x_N)$$

Bayes' rule

$$= \operatorname*{argmax}_{\mu,\sigma^2} p(\mu,\sigma^2)p(x_1,\cdots x_N|\mu,\sigma^2)$$

conditional independence

$$= \operatorname*{argmax}_{\mu,\sigma^2} p(\mu,\sigma^2)\prod_{n=1}^{N} p(x_n|\mu,\sigma^2)$$

Because log is monotonic

$$= \operatorname*{argmax}_{\mu,\sigma^2} \log\left(p(\mu,\sigma^2)\prod_{n=1}^{N} p(x_n|\mu,\sigma^2)\right)$$

$$\log(ab) = \log a + \log b$$

$$= \operatorname*{argmax}_{\mu,\sigma^2} \log p(\mu,\sigma^2) + \sum_{n=1}^{N} \log p(x_n|\mu,\sigma^2)$$

$$= \operatorname*{argmax}_{\mu,\sigma^2} \log \frac{1}{\sigma}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} - \frac{2\beta+\gamma(\delta-\mu)^2}{2\sigma^2} - \sum_{n=1}^{N}\left[\frac{(x_n-\mu)^2}{2\sigma^2} + \log\sqrt{2\pi\sigma^2}\right]$$

MAP estimation of μ, σ^2 given training data D = $\{x_1,\ldots,x_N\}$

- From previous slide:

$$(\hat{\mu}, \hat{\sigma}^2) = \underset{\mu,\sigma^2}{\operatorname{argmax}} \log \frac{1}{\sigma}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} - \frac{2\beta + \gamma(\delta-\mu)^2}{2\sigma^2} - \sum_{n=1}^{N}\left[\frac{(x_n-\mu)^2}{2\sigma^2} + \log\sqrt{2\pi\sigma^2}\right]$$

- Differentiating μ and setting to 0,

$$\hat{\mu} = \frac{\sum_{n=1}^{N} x_n + \gamma\delta}{N+\gamma} = \frac{N\mu_{ML} + \gamma\delta}{N+\gamma}$$

see uploaded Detailed derivations

  - Tradeoff between ML estimate μ_{ML} and prior mean δ

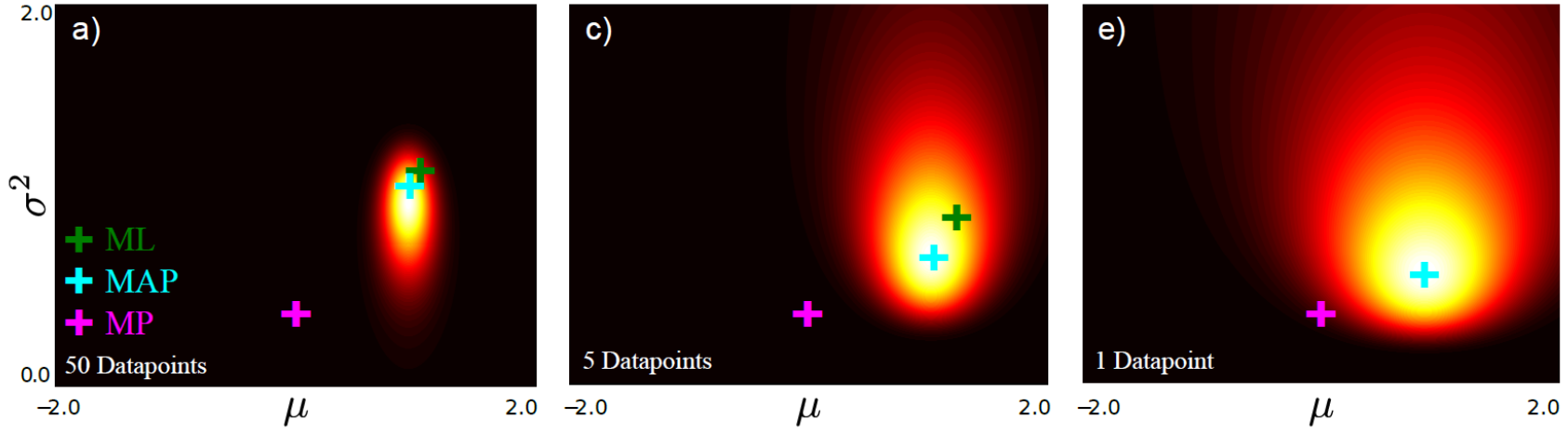
- Differentiating σ^2 and setting to 0,

$$\hat{\sigma}^2 = \frac{\sum_{n=1}^{N}(x_n-\hat{\mu})^2 + 2\beta + \gamma(\delta-\hat{\mu})^2}{N+3+2\alpha}$$

$$\overset{N\to\infty}{=} \frac{N\hat{\sigma}^2_{ML} + 2\beta + \gamma(\delta-\hat{\mu})^2}{N+3+2\alpha} = \hat{\sigma}^2_{ML}$$

see uploaded Detailed derivations

Fitting normal distribution: MAP



50 data points

5 data points

1 data points

Predicting Future Gaussian Samples

- Suppose we are given training data $D = \{x_1, \cdots, x_N\}$
- Let x^* be outcome of next Gaussian sample. What is the pdf of x^*?
- Strategy 1 (ML estimation): $p(x^*) = \mathcal{N}(\mu_{ML}, \sigma^2_{ML})$
- Strategy 2 (MAP estimation): $p(x^*) = \mathcal{N}(\mu_{MAP}, \sigma^2_{MAP})$
- Strategy 3 (Posterior Predictive Distribution)
  - Problem with ML/MAP strategies is that they are not quite optimal
  - Above question says "Suppose we are given D, what is the pdf of x^*", so the pdf we want to evaluate is $p(x^*|D)$, which is in fact not even Gaussian

Posterior Predictive Distribution

- Procedure same as dirichlet-multinomial and beta-bernoulli, but a bit tedious. Please refer to murphy-2007.pdf, which I will upload

- Predict new sample x^* given $x_{1:N}$:

$$p(x^*|x_{1:N}) = \kappa(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}, x^*)$$

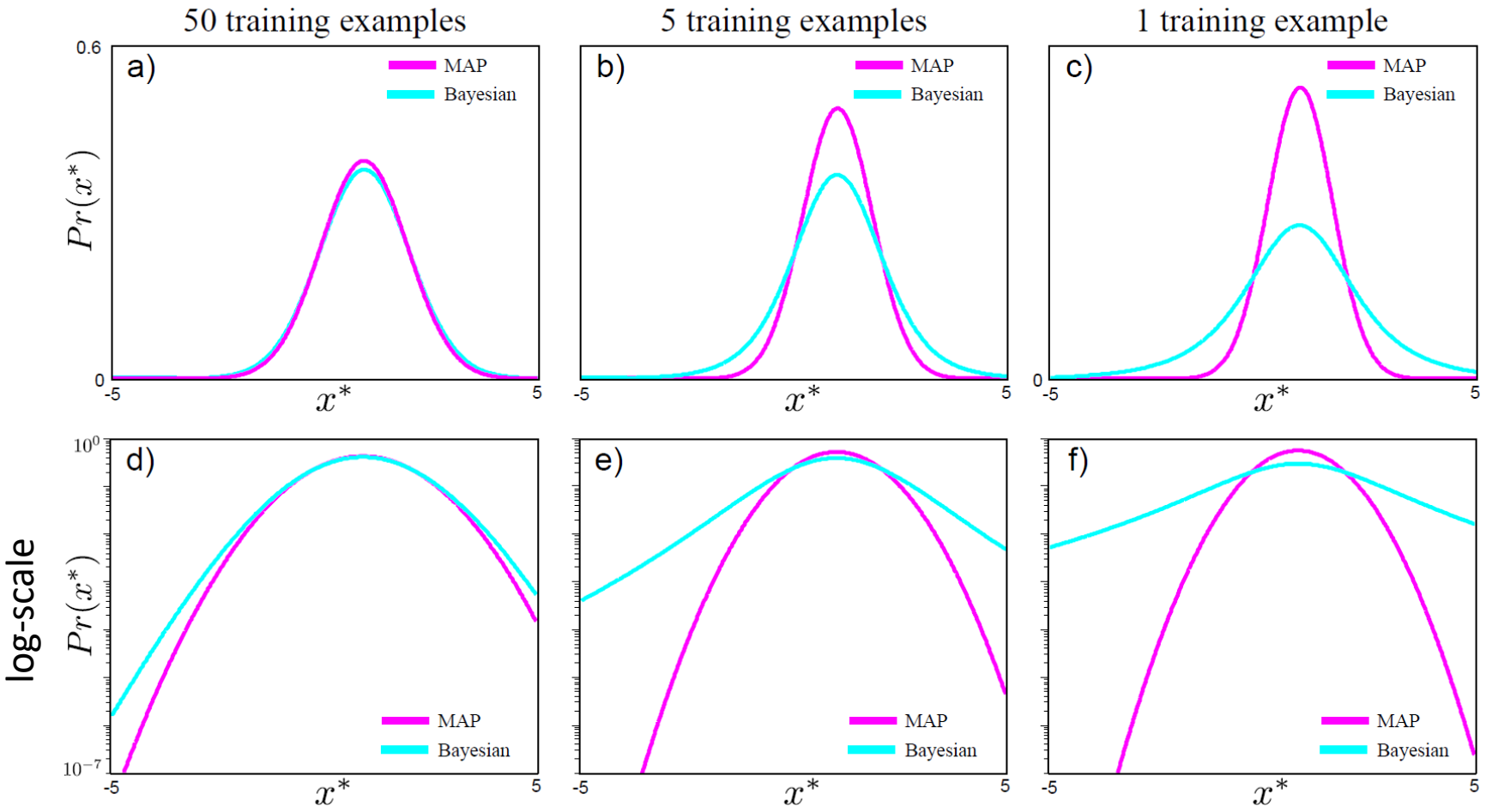
- $\kappa(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}, x^*) = \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\tilde{\gamma}}\tilde{\beta}^{\tilde{\alpha}}}{\sqrt{\dot{\gamma}}\dot{\beta}^{\dot{\alpha}}} \frac{\Gamma(\tilde{\alpha})}{\Gamma(\dot{\alpha})}$, where

$$\dot{\alpha} = \tilde{\alpha} + 1/2, \;\; \dot{\gamma} = \tilde{\gamma} + 1, \;\; \dot{\beta} = \frac{x^{*2}}{2} + \tilde{\beta} + \frac{\tilde{\gamma}\tilde{\delta}^2}{2} - \frac{(x^* + \tilde{\gamma}\tilde{\delta})^2}{2(\tilde{\gamma}+1)}$$

$$\tilde{\alpha} = \alpha + N/2, \;\; \tilde{\gamma} = \gamma + N, \;\; \tilde{\delta} = \frac{N\mu_{ML} + \gamma\delta}{\gamma + N}$$

$$\tilde{\beta} = \frac{\sum_n x_n^2}{2} + \beta + \frac{\gamma\delta^2}{2} - \frac{(\sum_n x_n + \gamma\delta)^2}{2(\gamma + N)}$$

Posterior Predictive Distribution vs MAP Plug-In



Naive Bayes Classifier

Naive Bayes Classifier I

- Let y = target class label and x = feature vector of length D
- Recall for generative classifier, we model $p(x, y \mid \boldsymbol{\theta}) = p(y \mid \boldsymbol{\theta})p(x \mid y, \boldsymbol{\theta})$
- To define classifier, need to define distributions $p(y \mid \boldsymbol{\theta})$ and $p(x \mid y, \boldsymbol{\theta})$
- Can split $\boldsymbol{\theta}$ into two parts: $\boldsymbol{\theta} = \{\lambda, \eta\}$
- λ are parameters of class prior: $p(y \mid \boldsymbol{\theta}) = p(y \mid \lambda)$
  - For classification, y is by definition discrete (and if y is single variable), then $p(y \mid \lambda)$ has to be a categorical distribution
  - Intuitively, class prior specifies frequency of a class (e.g., how often a chair appears in Facebook photos)
- η are parameters of feature likelihood: $p(x \mid y, \boldsymbol{\theta}) = p(x \mid y, \eta)$
  - Modeling aspect comes from specifying the feature likelihood
  - Intuitively, feature likelihood is saying that given the data point comes from class label c (e.g., chair), what is pdf of the features x?
  - Let x_j be j-th feature of x
  - Naive Bayes Classifier assumes features are conditionally independent given class label: $p(x \mid y = c, \eta) = \prod_{j=1}^{D} p(x_j \mid y = c, \eta) = \prod_{j=1}^{D} p(x_j \mid \eta_{jc})$, where η_{jc} are the parameters of the pdf of j-th feature assuming data point from class c

Naive Bayes Classifier II

- From previous slide: Naive Bayes assumes features are conditionally independent given class label: $p(x \mid y = c, \eta) = \prod_{j=1}^{D} p(x_j \mid y = c, \eta) = \prod_{j=1}^{D} p(x_j \mid \eta_{jc})$
- For example, suppose 1st feature x_1 is continuous and we model $p(x_1 \mid \eta_{1c})$ as Gaussian, then $\eta_{1c} = (\mu_{1c}, \sigma_{1c}^2)$
  - μ_{1c}, σ_{1c}^2 are mean & variance of 1st feature dimension if data comes from class c
  - Notice there is a mean & variance for each class c
- For example, suppose 2nd feature x_2 is binary and we model $p(x_2 \mid \eta_{2c})$ as bernoulli, then
  - η_{2c} is probability of observing head for 2nd feature dimension if the data comes from class c.
  - Notice there is an η for each class c
- There is an η for each class c and each feature j, so in total there are about $\mathcal{O}$(CD) parameters, where C = # target classes and D = # feature dimensions, $\mathcal{O}$ = "big-Oh" notation (think of $\mathcal{O}$ as "roughly equal to")
- "Naive" because features are probably not conditionally independent, but number of parameters scale linearly with C and D, so less likely to overfit
- Each feature dimension can be different distribution, e.g., feature 1 is Gaussian, feature 2 is binary, feature 3 is categorical, etc

Three Strategies for estimating Naive Bayes Parameters & Classifying New Sample

- Given training set D = $\{x_{ij}, y_i\}_{i=1:N, j=1:D}$, where x_{ij} is j-th feature of i-th data point (e.g., i-th photo) and y_i is the target class label (e.g., chair) of i-th data point
- Already specify generative model $p(x, y \mid \boldsymbol{\theta}) = p(y \mid \lambda)\, p(x \mid y, \eta)$
- Recall posterior $p(y \mid x, \boldsymbol{\theta}) \propto p(y \mid \lambda)\, p(x \mid y, \eta)$

Three Strategies for estimating Naive Bayes Parameters & Classifying New Sample

- Given training set D = $\{x_{ij}, y_i\}_{i=1:N, j=1:D}$, where x_{ij} is j-th feature of i-th data point (e.g., i-th photo) and y_i is the target class label (e.g., chair) of i-th data point
- Already specify generative model $p(x, y \mid \boldsymbol{\theta}) = p(y \mid \lambda)\, p(x \mid y, \eta)$
- Recall posterior $p(y \mid x, \boldsymbol{\theta}) \propto p(y \mid \lambda)\, p(x \mid y, \eta)$
- Strategy 1 (Maximum likelihood)
  - Step 1: Estimate $\boldsymbol{\theta}_{ML} = \text{argmax}_{\boldsymbol{\theta}}\, p(D \mid \boldsymbol{\theta})$
  - Step 2: To predict label $\tilde{y}$ of test data $\tilde{x}$, plug in $\boldsymbol{\theta}_{ML}$ into posterior $p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}) \propto p(\tilde{y} \mid \lambda_{ML})\, p(\tilde{x} \mid \tilde{y}, \eta_{ML})$ and find MAP estimate of $\tilde{y}$

Three Strategies for estimating Naive Bayes Parameters & Classifying New Sample

- Given training set D = $\{x_{ij}, y_i\}_{i=1:N, j=1:D}$, where x_{ij} is j-th feature of i-th data point (e.g., i-th photo) and y_i is the target class label (e.g., chair) of i-th data point
- Already specify generative model $p(x, y \mid \boldsymbol{\theta}) = p(y \mid \lambda)\, p(x \mid y, \eta)$
- Recall posterior $p(y \mid x, \boldsymbol{\theta}) \propto p(y \mid \lambda)\, p(x \mid y, \eta)$
- Strategy 1 (Maximum likelihood)
  - Step 1: Estimate $\boldsymbol{\theta}_{ML} = \text{argmax}_{\boldsymbol{\theta}}\, p(D \mid \boldsymbol{\theta})$
  - Step 2: To predict label $\tilde{y}$ of test data $\tilde{x}$, plug in $\boldsymbol{\theta}_{ML}$ into posterior $p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}) \propto p(\tilde{y} \mid \lambda_{ML})\, p(\tilde{x} \mid \tilde{y}, \eta_{ML})$ and find MAP estimate of $\tilde{y}$
- Strategy 2 (Maximum-A-Posteriori)
  - Step 1: Estimate $\boldsymbol{\theta}_{MAP} = \text{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta} \mid D)$
  - Step 2: To predict label $\tilde{y}$ of test data $\tilde{x}$, plug in $\boldsymbol{\theta}_{MAP}$ into posterior $p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}) \propto p(\tilde{y} \mid \lambda_{MAP})\, p(\tilde{x} \mid \tilde{y}, \eta_{MAP})$ and find MAP estimate of $\tilde{y}$

Three Strategies for estimating Naive Bayes Parameters & Classifying New Sample

- Given training set D = $\{x_{ij}, y_i\}_{i=1:N, j=1:D}$, where x_{ij} is j-th feature of i-th data point (e.g., i-th photo) and y_i is the target class label (e.g., chair) of i-th data point
- Already specify generative model $p(x, y \mid \boldsymbol{\theta}) = p(y \mid \lambda)\, p(x \mid y, \eta)$
- Recall posterior $p(y \mid x, \boldsymbol{\theta}) \propto p(y \mid \lambda)\, p(x \mid y, \eta)$
- Strategy 1 (Maximum likelihood)
  - Step 1: Estimate $\boldsymbol{\theta}_{ML} = \text{argmax}_{\boldsymbol{\theta}}\, p(D \mid \boldsymbol{\theta})$
  - Step 2: To predict label $\tilde{y}$ of test data $\tilde{x}$, plug in $\boldsymbol{\theta}_{ML}$ into posterior $p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}) \propto p(\tilde{y} \mid \lambda_{ML})\, p(\tilde{x} \mid \tilde{y}, \eta_{ML})$ and find MAP estimate of $\tilde{y}$
- Strategy 2 (Maximum-A-Posteriori)
  - Step 1: Estimate $\boldsymbol{\theta}_{MAP} = \text{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta} \mid D)$
  - Step 2: To predict label $\tilde{y}$ of test data $\tilde{x}$, plug in $\boldsymbol{\theta}_{MAP}$ into posterior $p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}) \propto p(\tilde{y} \mid \lambda_{MAP})\, p(\tilde{x} \mid \tilde{y}, \eta_{MAP})$ and find MAP estimate of $\tilde{y}$
- Strategy 3 (Posterior Predictive / Bayesian Model Averaging)
  - To predict label $\tilde{y}$ of test data $\tilde{x}$, compute posterior $p(\tilde{y} \mid \tilde{x}, D)$
  - Notice that posterior does not contain $\boldsymbol{\theta}$

Strategy 2: λ^{MAP}, η^{MAP} Can Be Estimated Separately

- Let's focus on MAP: $\theta_{MAP} = \text{argmax}_\theta \, p(\theta|D) = \text{argmax}_{\lambda,\eta} \, p(\lambda, \eta|D)$
- Recall $D = \{x_i, y_i\}_{i=1:N} = \{x_{1:N}, y_{1:N}\}$

$$p(\lambda, \eta|D) = \frac{p(\lambda, \eta)p(D|\lambda, \eta)}{p(D)}$$

Bayes' rule: $p(A|B) = \frac{p(A)p(B \mid A)}{p(B)}$

$$= \frac{p(\lambda, \eta)p(x_{1:N}, y_{1:N}|\lambda, \eta)}{p(x_{1:N}, y_{1:N})}$$

Plug in D = {$x_{1:N}$, $y_{1:N}$}

$$= \frac{p(\lambda)p(\eta)p(y_{1:N}|\lambda, \eta)p(x_{1:N}|y_{1:N}, \lambda, \eta)}{p(y_{1:N})p(x_{1:N}|y_{1:N})}$$

λ, η independent +
p(A, B|C) = p(B|C) p(A|B, C)

$$= \frac{p(\lambda)p(\eta)p(y_{1:N}|\lambda)p(x_{1:N}|y_{1:N}, \eta)}{p(y_{1:N})p(x_{1:N}|y_{1:N})}$$

Apply conditional independence relations

$$= \left(\frac{p(\lambda)p(y_{1:N}|\lambda)}{p(y_{1:N})}\right)\left(\frac{p(\eta)p(x_{1:N}|y_{1:N}, \eta)}{p(x_{1:N}|y_{1:N})}\right)$$

Re-grouping Terms

$$= p(\lambda|y_{1:N})\left(\frac{p(\eta)p(x_{1:N}|y_{1:N}, \eta)}{p(x_{1:N}|y_{1:N})}\right)$$

Bayes' rule:

$$= \text{can optimize } \lambda \text{ and } \eta \text{ separately}$$

Strategy 2 (Step 1): MAP Estimation of λ

- For $\text{argmax}_\lambda \, p(\lambda|y_{1:N})$
  - Assuming 2 classes (class 1 is "tail", class 2 is "head") and beta prior $\text{Beta}(a, b)$ $\implies$ exactly same as beta-binomial model in previous lecture, so apply MAP estimate from previous lecture

$$\lambda^{MAP} = \frac{N_{head} + a - 1}{N + a + b - 2},$$

    where N_{head} is number of data samples in class 2
  - Assuming >2 classes and dirichlet prior $\text{Dir}(\alpha_1, \cdots, \alpha_C)$ $\implies$ exactly same as dirichlet-multinomial model in previous lecture, so we apply MAP estimate from previous lecture

$$\lambda_c^{MAP} = \frac{N_c + \alpha_c - 1}{N + \sum_{c=1}^{C} \alpha_c - C},$$

    where N_c is number of data samples in class c

Strategy 2 (Step 1): MAP Estimation of η

- For η, observe we can optimize η_{jc} for each dimension j and class c separately

$$\frac{p(\eta)p(x_{1:N}|y_{1:N},\eta)}{p(x_{1:N}|y_{1:N})} = \prod_{j=1}^{D}\prod_{c=1}^{C}\frac{p(\eta_{jc})p(x_{i\in c,j}|\eta_{jc})}{p(x_{i\in c,j})}$$

- Each dimension j is independent because of naïve Bayes
- Knowing what chairs look like do not tell us what cows look like, so features from chair and cow photos are independent
- Priors for dimension j and class c are independent

where $x_{i\in c,j}$ indicates the j-th feature of all data points belonging to class c

Strategy 2 (Step 1): MAP Estimation of η

- For η, observe we can optimize η_{jc} for each dimension j and class c separately

$$\frac{p(\eta)p(x_{1:N}|y_{1:N},\eta)}{p(x_{1:N}|y_{1:N})} = \prod_{j=1}^{D}\prod_{c=1}^{C}\frac{p(\eta_{jc})p(x_{i\in c,j}|\eta_{jc})}{p(x_{i\in c,j})}$$

- Each dimension j is independent because of naïve Bayes
- Knowing what chairs look like do not tell us what cows look like, so features from chair and cow photos are independent
- Priors for dimension j and class c are independent

where $x_{i\in c,j}$ indicates the j-th feature of all data points belonging to class c

Strategy 2 (Step 1): MAP Estimation of η

- For η, observe we can optimize η_{jc} for each dimension j and class c separately

$$\frac{p(\eta)p(x_{1:N}|y_{1:N},\eta)}{p(x_{1:N}|y_{1:N})} = \prod_{j=1}^{D}\prod_{c=1}^{C}\frac{p(\eta_{jc})p(x_{i\in c,j}|\eta_{jc})}{p(x_{i\in c,j})}$$

- Each dimension j is independent because of naïve Bayes
- Knowing what chairs look like do not tell us what cows look like, so features from chair and cow photos are independent
- Priors for dimension j and class c are independent

where $x_{i\in c,j}$ indicates the j-th feature of all data points belonging to class c

Strategy 2 (Step 1): MAP Estimation of η

- For η, observe we can optimize η_{jc} for each dimension j and class c separately

$$\frac{p(\eta)p(x_{1:N}|y_{1:N},\eta)}{p(x_{1:N}|y_{1:N})} = \prod_{j=1}^{D}\prod_{c=1}^{C}\frac{p(\eta_{jc})p(x_{i\in c,j}|\eta_{jc})}{p(x_{i\in c,j})}$$

$$= \prod_{j=1}^{D}\prod_{c=1}^{C} p(\eta_{jc}|x_{i\in c,j})$$

Bayes' rule

- Each dimension j is independent because of naïve Bayes
- Knowing what chairs look like do not tell us what cows look like, so features from chair and cow photos are independent
- Priors for dimension j and class c are independent

where $x_{i\in c,j}$ indicates the j-th feature of all data points belonging to class c

- For every dimension j and every class c, we perform MAP estimation of η_{jc} using posterior $p(\eta_{jc}|x_{i\in c,j})$
  - For example, suppose feature 1 is Gaussian, to estimate $\eta_{12} = (\mu_{12}, \sigma^2_{12})$ (mean and variance of 1st feature of 2nd class), we apply formula from univariate Gaussian MAP lecture notes to 1st feature of all data samples coming from target class 2
  - For example, suppose feature 2 is binary, to estimate η_{23} (probability of observing head for 2nd feature of 3rd class), we apply formula from beta-binomial MAP lecture notes to 2nd feature of all data samples coming from target class 3

Strategy 2 (Step 2): Predicting Target Class of Test Data $\tilde{x}$ using MAP Estimates of λ, η

- Once we estimate λ^{MAP} (class prior parameters) and η^{MAP} (feature likelihood parameters) from training data, to predict target class label $\tilde{y}$ of test data $\tilde{x}$, we compute for each class c:

$$p(\tilde{y} = c|\tilde{x}, \lambda^{MAP}, \eta^{MAP})$$
$$\propto p(\tilde{y} = c|\lambda^{MAP}) p(\tilde{x}|\tilde{y} = c, \eta^{MAP})$$
$$= \lambda_c^{MAP} \prod_{j=1}^{D} p(\tilde{x}_j|\eta_{jc}^{MAP})$$

Recall posterior can be split into class prior & feature likelihood

Naïve Bayes Assumption

Strategy 2 (Step 2): Predicting Target Class of Test Data $\tilde{x}$ using MAP Estimates of λ, η

- Once we estimate λ^{MAP} (class prior parameters) and η^{MAP} (feature likelihood parameters) from training data, to predict target class label $\tilde{y}$ of test data $\tilde{x}$, we compute for each class c:

$$p(\tilde{y} = c|\tilde{x}, \lambda^{MAP}, \eta^{MAP})$$
$$\propto p(\tilde{y} = c|\lambda^{MAP}) p(\tilde{x}|\tilde{y} = c, \eta^{MAP})$$
$$= \lambda_c^{MAP} \prod_{j=1}^{D} p(\tilde{x}_j|\eta_{jc}^{MAP})$$

Recall posterior can be split into class prior & feature likelihood

Naïve Bayes Assumption

Strategy 2 (Step 2): Predicting Target Class of Test Data $\tilde{x}$ using MAP Estimates of λ, η

- Once we estimate λ^{MAP} (class prior parameters) and η^{MAP} (feature likelihood parameters) from training data, to predict target class label $\tilde{y}$ of test data $\tilde{x}$, we compute for each class c:

$$p(\tilde{y} = c|\tilde{x}, \lambda^{MAP}, \eta^{MAP})$$
$$\propto p(\tilde{y} = c|\lambda^{MAP}) p(\tilde{x}|\tilde{y} = c, \eta^{MAP})$$
$$= \lambda_c^{MAP} \prod_{j=1}^{D} p(\tilde{x}_j|\eta_{jc}^{MAP})$$

Recall posterior can be split into class prior & feature likelihood

Naïve Bayes Assumption

- The MAP estimate of class c is class with highest posterior $p(\tilde{y} = c|\tilde{x}, \lambda^{MAP}, \eta^{MAP})$
- In practice, for numerical stability, we compute

$$\log \lambda_c^{MAP} \prod_{j=1}^{D} p(\tilde{x}_j|\eta_{jc}^{MAP})$$

Strategy 2 (Step 2): Predicting Target Class of Test Data $\tilde{x}$ using MAP Estimates of λ, η

- Once we estimate λ^{MAP} (class prior parameters) and η^{MAP} (feature likelihood parameters) from training data, to predict target class label $\tilde{y}$ of test data $\tilde{x}$, we compute for each class c:

$$p(\tilde{y} = c|\tilde{x}, \lambda^{MAP}, \eta^{MAP})$$
$$\propto p(\tilde{y} = c|\lambda^{MAP}) p(\tilde{x}|\tilde{y} = c, \eta^{MAP})$$
$$= \lambda_c^{MAP} \prod_{j=1}^{D} p(\tilde{x}_j|\eta_{jc}^{MAP})$$

Recall posterior can be split into class prior & feature likelihood

Naïve Bayes Assumption

- The MAP estimate of class c is class with highest posterior $p(\tilde{y} = c|\tilde{x}, \lambda^{MAP}, \eta^{MAP})$
- In practice, for numerical stability, we compute

$$\log \lambda_c^{MAP} \prod_{j=1}^{D} p(\tilde{x}_j|\eta_{jc}^{MAP}) = \log \lambda_c^{MAP} + \sum_{j=1}^{D} \log p(\tilde{x}_j|\eta_{jc}^{MAP})$$

log ab = log a + log b

- For strategy 1, instead of using MAP estimates $\lambda^{MAP}, \eta^{MAP}$, use ML estimates λ^{ML}, η^{ML}

Naïve Bayes Example

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & ML for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & ML for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1 | \tilde{x}, \lambda^{ML}, \eta^{ML})$$

$$\propto \log p(\tilde{y} = 1 | \lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j | \eta_{j1}^{ML})$$

Formula from previous slide: c = 1

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1 | \eta_{11}^{ML}) + \log p(\tilde{x}_2 = 0 | \eta_{21}^{ML})$$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & ML for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1 | \tilde{x}, \lambda^{ML}, \eta^{ML})$$

$$\propto \log p(\tilde{y} = 1 | \lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j | \eta_{j1}^{ML})$$

Formula from previous slide: c = 1

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1 | \eta_{11}^{ML}) + \log p(\tilde{x}_2 = 0 | \eta_{21}^{ML})$$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & ML for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1 | \tilde{x}, \lambda^{ML}, \eta^{ML})$$

$$\propto \log p(\tilde{y} = 1 | \lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j | \eta_{j1}^{ML})$$

Formula from previous slide: c = 1

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1 | \eta_{11}^{ML}) + \log p(\tilde{x}_2 = 0 | \eta_{21}^{ML})$$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & ML for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1 | \tilde{x}, \lambda^{ML}, \eta^{ML})$$

$$\propto \log p(\tilde{y} = 1 | \lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j | \eta_{j1}^{ML})$$

Formula from previous slide: c = 1

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1 | \eta_{11}^{ML}) + \log p(\tilde{x}_2 = 0 | \eta_{21}^{ML})$$

$$= \log 1/3 + \log 0 + \log(1 - 0) = -\infty$$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & ML for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1 | \tilde{x}, \lambda^{ML}, \eta^{ML})$$

$$\propto \log p(\tilde{y} = 1 | \lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j | \eta_{j1}^{ML})$$

Formula from previous slide: c = 1

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1 | \eta_{11}^{ML}) + \log p(\tilde{x}_2 = 0 | \eta_{21}^{ML})$$

$$= \log 1/3 + \log 0 + \log(1 - 0) = -\infty$$

- Last lecture: ML estimate of probability of head $= \frac{N_1}{N}$, where N_1 = # heads & N = # coin tosses
  - Class 1, feature 1: N = 1 (since class 1 appears once), $N_1 = 0$ (since among class 1 training data, feature 1 is always 0), so $\eta_{11}^{ML} = \frac{0}{1} = 0$
  - Class 1, feature 2: N = 1 (since class 1 appears once), $N_1 = 0$ (since among class 1 training data, feature 2 is always 0) so $\eta_{21}^{ML} = \frac{0}{1} = 0$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & ML for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 2|\tilde{x}, \lambda^{ML}, \eta^{ML})$$

$$\propto \log p(\tilde{y} = 2|\lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j|\eta_{j2}^{ML})$$

Formula from previous slide: c = 2

$$= \log \lambda^{ML} + \log p(\tilde{x}_1 = 1|\eta_{12}^{ML}) + \log p(\tilde{x}_2 = 0|\eta_{22}^{ML})$$

$$= \log 2/3 + \log 1 + \log(1 - 1/2) = -1.10$$

- Last lecture: ML estimate of probability of head $= \frac{N_1}{N}$, where N_1 = # heads & N = # coin tosses
  - Class 2, feature 1: N = 2 (since class 2 appears twice), $N_1 = 2$ (since among class 2 training data, feature 1 is 1 twice), so $\eta_{12}^{ML} = \frac{2}{2} = 1$
  - Class 2, feature 2: N = 2 (since class 2 appears twice), $N_1 = 1$ (since among class 2 training data, feature 2 is 1 once), so $\eta_{22}^{ML} = \frac{1}{2} = 1/2$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & ML for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1|\tilde{x}, \lambda^{ML}, \eta^{ML})$$

$$\propto \log p(\tilde{y} = 1|\lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j|\eta_{j1}^{ML})$$

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1|\eta_{11}^{ML}) + \log p(\tilde{x}_2 = 0|\eta_{21}^{ML})$$

$$= \log 1/3 + \log 0 + \log(1 - 0) = -\infty$$

$$\log p(\tilde{y} = 2|\tilde{x}, \lambda^{ML}, \eta^{ML})$$

$$\propto \log p(\tilde{y} = 2|\lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j|\eta_{j2}^{ML})$$

$$= \log \lambda^{ML} + \log p(\tilde{x}_1 = 1|\eta_{12}^{ML}) + \log p(\tilde{x}_2 = 0|\eta_{22}^{ML})$$

$$= \log 2/3 + \log 1 + \log(1 - 1/2) = -1.10$$

- -1.10 is bigger than $-\inf$, so predict $\tilde{y} = 2$

Strategy 3: Posterior Predictive

- Recall $D = \{x_i, y_i\}_{i=1:N} = \{x_{1:N}, y_{1:N}\}$
- For posterior predictive estimation of test data $(\tilde{x}, \tilde{y})$, instead of considering $p(\tilde{y}|\tilde{x}, \theta)$ (like in MAP estimation), we consider different posterior by essentially skipping "middle man" θ

$$p(\tilde{y}|\tilde{x}, D) = \frac{p(\tilde{x}, \tilde{y}|D)}{p(\tilde{x}|D)}$$

$p(\tilde{x}, \tilde{y} \mid D) = p(\tilde{x} \mid D)\, p(\tilde{y} \mid \tilde{x}, D)$

$$\propto p(\tilde{x}, \tilde{y}|D)$$

$p(\tilde{x} \mid D)$ does not depend on $\tilde{y}$

$$= p(\tilde{y}|D)p(\tilde{x}|\tilde{y}, D)$$

$p(A, B \mid C) = p(A \mid C)p(B|A, C)$

$$= p(\tilde{y}|x_{1:N}, y_{1:N})p(\tilde{x}|\tilde{y}, x_{1:N}, y_{1:N})$$

Plug in $D = \{x_{1:N}, y_{1:N}\}$

$$= p(\tilde{y}|y_{1:N})p(\tilde{x}|\tilde{y}, x_{1:N}, y_{1:N})$$

Apply independence relations

$$=$$ Evaluate two terms separately

- First term $p(\tilde{y}|y_{1:N})$ is just posterior predictive distribution for beta-binomial (2 classes) or dirichlet-multinomial (>2 classes) model (see last week notes)

Strategy 3: Posterior Predictive

- For second term $p(\tilde{x}|\tilde{y}, x_{1:N}, y_{1:N})$, observe that we can again compute independently for each dimension j and class c

  - Each dimension j is independent because of naïve Bayes

$$p(\tilde{x}|\tilde{y} = c, x_{1:N}, y_{1:N}) = \prod_{j=1}^{D} p(\tilde{x}_j|x_{i\in c,j}, \tilde{y} = c),$$

  where $x_{i\in c,j}$ indicates the j-th feature of all training data points belonging to class c

Strategy 3: Posterior Predictive

- For second term $p(\tilde{x}|\tilde{y}, x_{1:N}, y_{1:N})$, observe that we can again compute independently for each dimension j and class c

$$p(\tilde{x}|\tilde{y} = c, x_{1:N}, y_{1:N}) = \prod_{j=1}^{D} p(\tilde{x}_j|x_{i \in c,j}, \tilde{y} = c),$$

where $x_{i \in c,j}$ indicates the j-th feature of all training data points belonging to class c

- Each dimension j is independent because of naïve Bayes
- Knowing what chairs look like do not tell us what cows look like, so features from chair and cow photos are independent

Strategy 3: Posterior Predictive

- For second term $p(\tilde{x}|\tilde{y}, x_{1:N}, y_{1:N})$, observe that we can again compute independently for each dimension j and class c

$$p(\tilde{x}|\tilde{y} = c, x_{1:N}, y_{1:N}) = \prod_{j=1}^{D} p(\tilde{x}_j|x_{i\in c,j}, \tilde{y} = c),$$

  - Each dimension j is independent because of naïve Bayes
  - Knowing what chairs look like do not tell us what cows look like, so features from chair and cow photos are independent

  where $x_{i\in c,j}$ indicates the j-th feature of all training data points belonging to class c

- For every dimension j and every class c, we compute posterior predictive distribution of test feature $p(\tilde{x}_j|x_{i\in c,j}, \tilde{y} = c)$

  - For example, suppose feature 1 is Gaussian. Then for class 2, we compute the pdf of observing the first feature of the test sample ($\tilde{x}_1$) using univariate Gaussian posterior predictive distribution lecture notes based on the 1st feature of all data samples coming from target class 2
  - For example, suppose feature 2 is binary. Then for class 3, we compute the probability of observing the second feature of the test sample ($\tilde{x}_2$) using beta-binomial posterior predictive distribution lecture notes based on the 2nd feature of all data samples coming from target class 3

Predicting Target Class of Test Data $\tilde{x}$ Using Posterior Predictive Distribution

- To summarize, to predict target class label $\tilde{y}$ of test data $\tilde{x}$, we compute for each class c:

$$p(\tilde{y} = c|\tilde{x}, D) \propto p(\tilde{y} = c|y_{1:N}) \prod_{j=1}^{D} p(\tilde{x}_j|x_{i \in c,j}, \tilde{y} = c)$$

- The MAP estimate of class c is the class with highest posterior $p(\tilde{y} = c|\tilde{x}, D)$
- In practice, for numerical stability, we compute

$$\log p(\tilde{y} = c|\tilde{x}, D) \propto \log p(\tilde{y} = c|y_{1:N}) + \sum_{j=1}^{D} \log p(\tilde{x}_j|x_{i \in c,j}, \tilde{y} = c)$$

log ab = log a + log b

- Furthermore, we do not need to same estimation scheme within same problem, e.g., in programming assignment, we will use ML estimate for class prior, and posterior predictive for binary feature likelihood, so above becomes

$$\log p(\tilde{y} = c|\tilde{x}, D) \propto \log p(\tilde{y} = c|\lambda^{ML}) + \sum_{j=1}^{D} \log p(\tilde{x}_j|x_{i \in c,j}, \tilde{y} = c)$$

Naïve Bayes Example

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y} = ?$

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & posterior predictive with prior Beta(2, 2) for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1|\tilde{x}, D)$$

$$\propto \log p(\tilde{y} = 1|\lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j|x_{i\in 1,j}, \tilde{y} = 1)$$

Formula from previous slide: c = 1

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1|x_{i\in 1,1}, \tilde{y} = 1) + \log p(\tilde{x}_2 = 0|x_{i\in 1,2}, \tilde{y} = 1)$$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y} = ?$

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & posterior predictive with prior Beta(2, 2) for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1 | \tilde{x}, D)$$

$$\propto \log p(\tilde{y} = 1 | \lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j | x_{i \in 1,j}, \tilde{y} = 1)$$

Formula from previous slide: c = 1

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1 | x_{i \in 1,1}, \tilde{y} = 1) + \log p(\tilde{x}_2 = 0 | x_{i \in 1,2}, \tilde{y} = 1)$$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & posterior predictive with prior Beta(2, 2) for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1|\tilde{x}, D)$$

$$\propto \log p(\tilde{y} = 1|\lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j|x_{i\in 1,j}, \tilde{y} = 1)$$

Formula from previous slide: c = 1

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1|x_{i\in 1,1}, \tilde{y} = 1) + \log p(\tilde{x}_2 = 0|x_{i\in 1,2}, \tilde{y} = 1)$$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & posterior predictive with prior Beta(2, 2) for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1|\tilde{x}, D)$$

$$\propto \log p(\tilde{y} = 1|\lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j|x_{i\in 1,j}, \tilde{y} = 1)$$

Formula from previous slide: c = 1

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1|x_{i\in 1,1}, \tilde{y} = 1) + \log p(\tilde{x}_2 = 0|x_{i\in 1,2}, \tilde{y} = 1)$$

$$= \log 1/3 + \log 2/5 + \log(1 - 2/5) = -2.53$$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & posterior predictive with prior Beta(2, 2) for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1 | \tilde{x}, D)$$

$$\propto \log p(\tilde{y} = 1 | \lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j | x_{i \in 1,j}, \tilde{y} = 1)$$

Formula from previous slide: c = 1

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1 | x_{i \in 1,1}, \tilde{y} = 1) + \log p(\tilde{x}_2 = 0 | x_{i \in 1,2}, \tilde{y} = 1)$$

$$= \log 1/3 + \log 2/5 + \log(1 - 2/5) = -2.53$$

- Last lecture: $p(\tilde{x} = 1 \mid D) = \frac{N_1 + a}{N + a + b} = \frac{N_1 + 2}{N + 4}$, where N_1 = # heads & N = # coin tosses
  - Class 1, feature 1: N = 1 (since class 1 appears once), $N_1 = 0$ (since among class 1 training data, feature 1 is always 0), so $p(\tilde{x}_1 = 1 \mid x_{i \in 1,1}, \tilde{y} = 1) = \frac{0+2}{1+4} = 2/5$
  - Class 1, feature 2: N = 1 (since class 1 appears once), $N_1 = 0$ (since among class 1 training data, feature 2 is always 0), so $p(\tilde{x}_2 = 1 \mid x_{i \in 1,2}, \tilde{y} = 1) = \frac{0+2}{1+4} = 2/5$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y}$ = ?

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & posterior predictive with prior Beta(2, 2) for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 2|\tilde{x}, D)$$

$$\propto \log p(\tilde{y} = 2|\lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j|x_{i\in 2,j}, \tilde{y} = 2)$$

Formula from previous slide: c = 2

$$= \log \lambda^{ML} + \log p(\tilde{x}_1 = 1|x_{i\in 2,1}, \tilde{y} = 2) + \log p(\tilde{x}_2 = 0|x_{i\in 2,2}, \tilde{y} = 2)$$

$$= \log 2/3 + \log 2/3 + \log(1 - 1/2) = -1.50$$

- Last lecture: $p(\tilde{x} = 1 \mid D) = \frac{N_1+a}{N+a+b} = \frac{N_1+2}{N+4}$, where N_1 = # heads & N = # coin tosses
  - Class 2, feature 1: N = 2 (since class 2 appears twice), $N_1 = 2$ (since among class 2 training data, feature 1 is 1 twice), so $p(\tilde{x}_1 = 1 \mid x_{i\in 2,1}, \tilde{y} = 2) = \frac{2+2}{2+4} = 2/3$
  - Class 2, feature 2: N = 2 (since class 2 appears twice), $N_1 = 1$ (since among class 2 training data, feature 2 is 1 once), so $p(\tilde{x}_2 = 1 \mid x_{i\in 2,2}, \tilde{y} = 2) = \frac{1+2}{2+4} = 1/2$

Feature 1	Feature 2	Class Label
$x_{11} = 1$	$x_{12} = 0$	$y_1 = 2$
$x_{21} = 1$	$x_{22} = 1$	$y_2 = 2$
$x_{31} = 0$	$x_{32} = 0$	$y_3 = 1$
$\tilde{x}_1 = 1$	$\tilde{x}_2 = 0$	$\tilde{y} = ?$

Training Data

Test Data

- Class 2 = head, Class 1 = tail
- Predict $\tilde{y}$ using ML for class prior & posterior predictive with prior Beta(2, 2) for feature likelihood

- $\lambda^{ML} = 2/3$ because 2 out of 3 training examples are class 2 (heads)

$$\log p(\tilde{y} = 1|\tilde{x}, D)$$

$$\propto \log p(\tilde{y} = 1|\lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j|x_{i\in 1,j}, \tilde{y} = 1)$$

$$= \log(1 - \lambda^{ML}) + \log p(\tilde{x}_1 = 1|x_{i\in 1,1}, \tilde{y} = 1) + \log p(\tilde{x}_2 = 0|x_{i\in 1,2}, \tilde{y} = 1)$$

$$= \log 1/3 + \log 2/5 + \log(1 - 2/5) = -2.53$$

$$\log p(\tilde{y} = 2|\tilde{x}, D)$$

$$\propto \log p(\tilde{y} = 2|\lambda^{ML}) + \sum_{j=1}^{2} \log p(\tilde{x}_j|x_{i\in 2,j}, \tilde{y} = 2)$$

$$= \log \lambda^{ML} + \log p(\tilde{x}_1 = 1|x_{i\in 2,1}, \tilde{y} = 2) + \log p(\tilde{x}_2 = 0|x_{i\in 2,2}, \tilde{y} = 2)$$

$$= \log 2/3 + \log 2/3 + \log(1 - 1/2) = -1.50$$

- -1.50 is bigger than -2.53, so predict $\tilde{y} = 2$

Summary

- Univariate Gaussian
  - ML, MAP, posterior predictive
- Naïve Bayes Classifier
  - Generative classifier: $p(x, y \mid \boldsymbol{\theta}) = p(y \mid \lambda)p(x \mid y, \eta)$
  - Features are independent given class labels
  - ML, MAP or posterior predictive strategies for estimating model parameters and classifying new test sample

Optional Reading

- Notes based on
  - SP Chapters 4.4 (free: www.computervisionmodels.com)
  - KM Chapter 3.5 (beware of typos)
- I won't test you on Big-Oh notation, but for brief introduction, see https://www.khanacademy.org/computing/computer-science/algorithms/asymptotic-notation/a/asymptotic-notation