ISTA 421/521
Introduction to Machine Learning

LECTURE NINETEEN: BAYESIAN CLASSIFICATION

COUNT MORRISON

clayton@sista.arizona.edu

Gould–Simpson 819

Phone 621–6609

31 - 1 October 2013

Main parametric modeling frameworks

- Minimizing a Loss function
  - Linear model
  - Linear least mean squares
- Maximum Likelihood
  - Probabilistic model of uncertainty (noise, error)
  - Maximize the likelihood w.r.t. parameters
  - Linear model with additive Gaussian noise
- Bayesian Approach
  - Treat parameters as random variables
  - Use Bayes Theorem to combine likelihood & prior to find posterior distribution

Approaches to Avoiding Over-fitting
i.e., how to achieve ***generalization***
- Regularization
- Cross Validation (estimating the gen error)
- Marginal Likelihood model selection

- Estimation Techniques (often used in Bayesian approaches)
  - Gradient methods (Widrow-Hoff (1st), Newton-Rhapson (2nd))
  - Laplace Approximation
  - Monte Carlo estimation of expectation; Metropolis-Hastings
- Classification (& Regression)
- Clustering
- Projection

predicting output

Main algorithmic families of Machine Learning

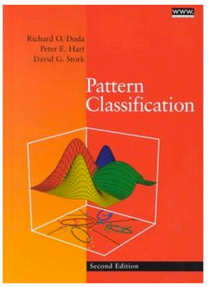
Classification

- *N* training objects, $\mathbf{x}_1$, …, $\mathbf{x}_N$
- Each $\mathbf{x}_i$ is a *D*-dimensional vector
- Each object has a label, t_n describing the class object *n* belongs to
  - Typically class label is expressed as an integer
  - Binary case:
    - $t_n = \{0,1\}$
    - $t_n = \{-1,1\}$
  - *C* classes:
    - $t_n = \{1, 2, \ldots, C\}$
- Task: predict the class t_{new} for an unseen object $\mathbf{x}_{new}$

Issues in Classification

Duda
Hart
Stork
2001

Ch 2 !



- Different domains have different problems.
  - Disease Diagnosis: How do we handle the uneven cost of making errors?
  - Text classification: How do we handle complex data objects like text?

- Two very general ML approaches to Classification:
  - Non-probabilistic – if all we care about is class assignment (often, just define decision boundary)
  - Probabilistic – permits measure of *confidence* in class assignment

Probabilistic Classifiers

$$P(T_{\text{new}} = c|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$$

$$0 \leq P(T_{\text{new}} = c|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) \leq 1$$

$$\sum_{c=1}^{C} P(T_{\text{new}} = c|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = 1$$

Note: assumes mutual exclusivity!

Disease classification:

$$P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = 0.6$$

versus

$$P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = 0.9$$

The Bayes Classifier

- Given a set of training points from *C* classes

$$P(T_{\text{new}} = c|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) =$$

$$P(T_{\text{new}} = c|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = \frac{p(\mathbf{x}_{\text{new}}|T_{\text{new}} = c, \mathbf{X}, \mathbf{t})P(T_{\text{new}} = c|\mathbf{X}, \mathbf{t})}{\sum_{c'=1}^{C} p(\mathbf{x}_{\text{new}}|T_{\text{new}} = c', \mathbf{X}, \mathbf{t})P(T_{\text{new}} = c'|\mathbf{X}, \mathbf{t})}$$

$$p(\mathbf{x}_{\text{new}}|T_{\text{new}} = c, \mathbf{X}, \mathbf{t}) \text{ and } P(T_{\text{new}} = c|\mathbf{X}, \mathbf{t})$$

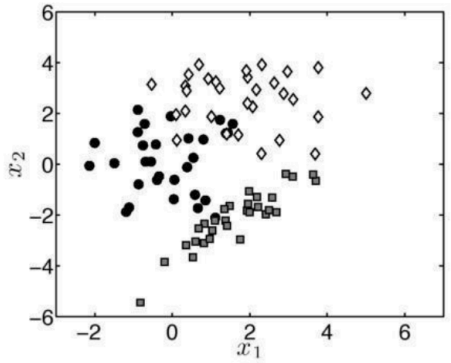
Need to define *C* class-conditional distributions
Usually these are the same type (but not necessarily)
Then find parameters (ML!)

Probability of *c* "before evidence"
Can account for uneven class sizes (can bias for or against a class)
Always: must be positive and $\sum_c P(T_{\text{new}} = c|\mathbf{X}, \mathbf{t}) = 1$

1. Uniform prior: $P(T_{\text{new}} = c|\mathbf{X}, \mathbf{t}) = \frac{1}{C}$
2. Class size prior: $P(T_{\text{new}} = c|\mathbf{X}, \mathbf{t}) = \frac{N_c}{N}$

The Bayes Classifier

- Example: Gaussian class-conditionals



$$\mathbf{x}_n = [x_{n1}, x_{n2}]^\mathsf{T}$$
$$t_n = \{1, 2, 3\}$$
$$p(\mathbf{x}_n | t_n = c, \mathbf{X}, \mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Next step: estimate $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

One possibility: A Bayesian approach…

$$p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | \mathbf{X}^c) = \frac{p(\mathbf{X}^c | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}{p(\mathbf{X}^c)}$$

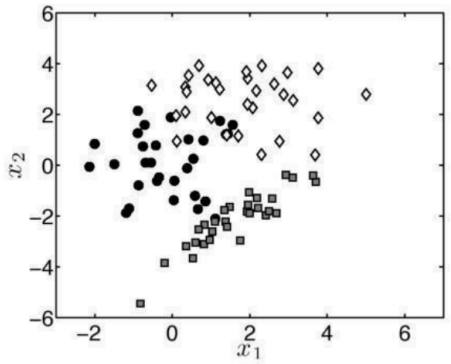
….then compute the likelihood of $\mathbf{x}_{\text{new}}$ by taking this expectation:

$$p(\mathbf{x}_{\text{new}} | T_{\text{new}} = c, \mathbf{X}, \mathbf{t}) = \mathbf{E}_{p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | \mathbf{X}^c)} \{ p(\mathbf{x}_{\text{new}} | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \}$$

This is most useful when there is little data and our estimates of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are uncertain

The Bayes Classifier

- Example: Gaussian class-conditionals



$$\mathbf{x}_n = [x_{n1}, x_{n2}]^\mathsf{T}$$
$$t_n = \{1, 2, 3\}$$
$$p(\mathbf{x}_n | t_n = c, \mathbf{X}, \mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Next step: estimate $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

Direct maximum likelihood estimates of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

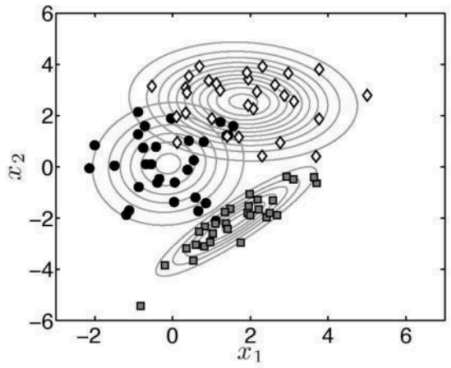
$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} \mathbf{x}_n$$
$$\boldsymbol{\Sigma}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} (\mathbf{x}_n - \boldsymbol{\mu}_c)(\mathbf{x}_n - \boldsymbol{\mu}_c)^\mathsf{T}$$

Summations are only for the data instances from the c^{th} class.

$$P(T_{\text{new}} = c | \mathbf{X}, \mathbf{t}) = \tfrac{1}{3}$$

The Bayes Classifier

- Example: Gaussian class-conditionals



$$\mathbf{x}_n = [x_{n1}, x_{n2}]^\mathsf{T}$$
$$t_n = \{1, 2, 3\}$$
$$p(\mathbf{x}_n | t_n = c, \mathbf{X}, \mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Next step: estimate $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

Direct maximum likelihood estimates of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

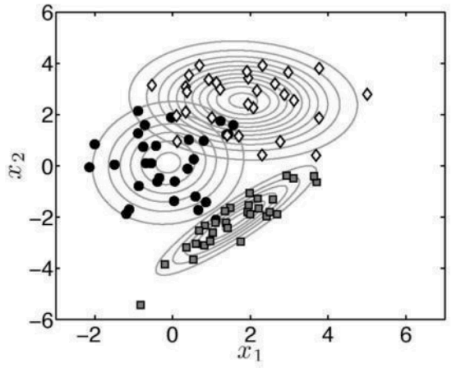
$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} \mathbf{x}_n$$
$$\boldsymbol{\Sigma}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} (\mathbf{x}_n - \boldsymbol{\mu}_c)(\mathbf{x}_n - \boldsymbol{\mu}_c)^\mathsf{T}$$

Summations are only for the data instances from the c^{th} class.

$P(T_{\text{new}} = c | \mathbf{X}, \mathbf{t}) = \frac{1}{3}$

The Bayes Classifier

- Example: Gaussian class-conditionals



$$\mathbf{x}_n = [x_{n1}, x_{n2}]^\mathsf{T}$$
$$t_n = \{1, 2, 3\}$$
$$p(\mathbf{x}_n | t_n = c, \mathbf{X}, \mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Next step: estimate $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

Making Predictions for $\mathbf{x}_{\text{new}} = [2, 0]^\mathsf{T}$

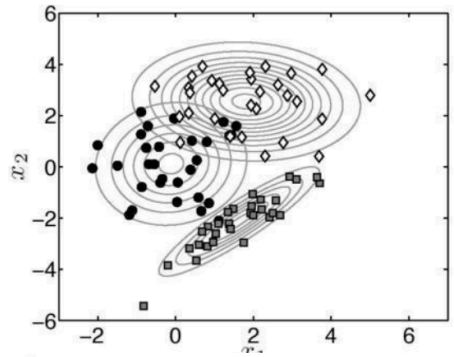
c	$p(\mathbf{x}_{\text{new}} \mid T_{\text{new}} = c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$	$P(T_{\text{new}} = c \mid \mathbf{X}, \mathbf{t})$	$p(\mathbf{x}_{\text{new}} \mid T_{\text{new}} = c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) P(T_{\text{new}} = c \mid \mathbf{X}, \mathbf{t})$	
1	0.0138	$\frac{1}{3}$	0.0046	0.6890
2	0.0061	$\frac{1}{3}$	0.0020	0.3024
3	0.0002	$\frac{1}{3}$	0.0001	0.0087

normalized

$$0.0046 + 0.0020 + 0.0001 = 0.0067$$

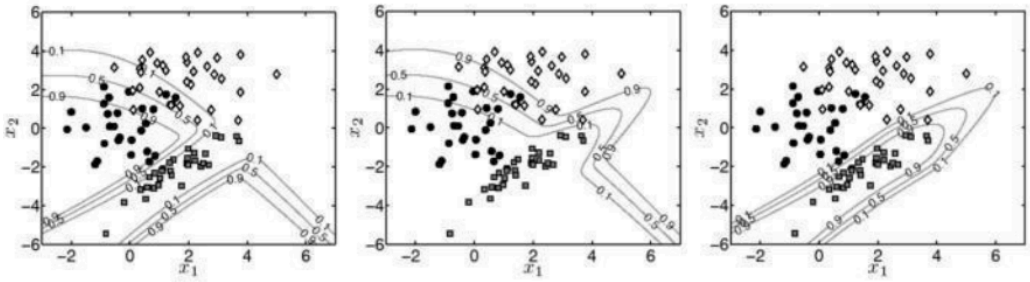
The Bayes Classifier

- Example: Gaussian class-conditionals



$$\mathbf{x}_n = [x_{n1}, x_{n2}]^\mathsf{T}$$
$$t_n = \{1, 2, 3\}$$
$$p(\mathbf{x}_n | t_n = c, \mathbf{X}, \mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Next step: estimate $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$



(a) $P(T_{\text{new}} = 1 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$ (b) $P(T_{\text{new}} = 2 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$ (c) $P(T_{\text{new}} = 3 | \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$

The Bayes Classifier

- Problem with Bayes Classifier: Growth of parameters to estimate as number of dimensions (D) increases.
- For Gaussian class-conditional Bayes:

$$D + D + \frac{D(D-1)}{2}$$

For the mean

For the covariance matrix (diagonally symmetrical)

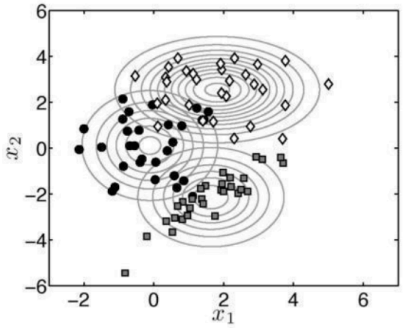
For 10 dimensions, 30 data points are not sufficient to estimate 65 parameters!

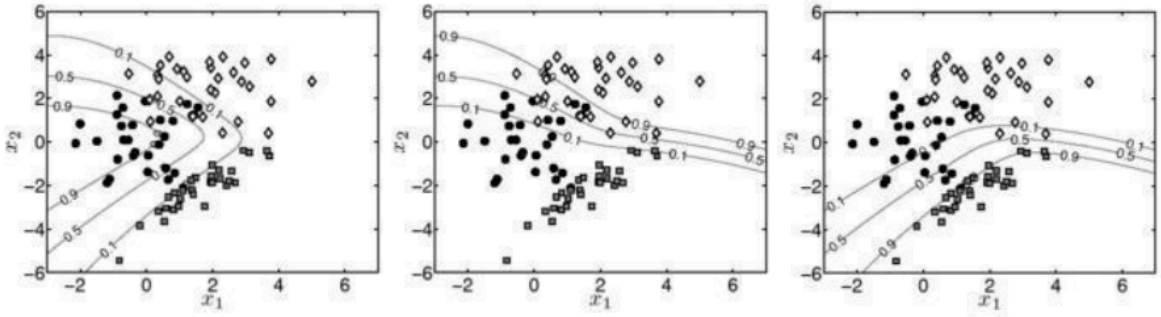
The key cost here is the covariance estimate.
We can dramatically simplify by assuming (class-conditional) ***independence***

Naïve Bayes Classifier

- The Gaussian class-conditional

$$p(\mathbf{x}_n|t_n = k, \mathbf{X}, \mathbf{t}) = \prod_{d=1}^{2} p(x_{nd}|t_n = k, \mathbf{X}, \mathbf{t})$$





(a) $P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$ (b) $P(T_{\text{new}} = 2|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$ (c) $P(T_{\text{new}} = 3|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$

Example Naïve Bayes Classifier

$$p(C|F_1, F_2, ..., F_k, C) = \frac{p(F_1|C)p(F_2|C)\cdots p(F_k|C)p(C)}{p(F_1, F_2, ..., F_k)}$$

$$p(F = x_i|C) = \alpha_{1c}^{[F=x_1]}\alpha_{2c}^{[F=x_2]}\cdots\alpha_{kc}^{[F=x_k]}$$

Example:

Categorical class-conditionals from nominal data: predicting ***play***

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Example Naïve Bayes Classifier

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/	5/
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5	14	14
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

$$p(F = x_i|C) = \alpha_{1c}^{[F=x_1]}\alpha_{2c}^{[F=x_2]}\cdots\alpha_{kc}^{[F=x_k]}$$

Example:

Categorical class-conditionals from nominal data: predicting ***play***

Example Naïve Bayes Classifier

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/	5/
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5	14	14
Rainy	3/9	2/5	Cool	3/9	1/5								

A new day:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

$$p(F = x_i|C) = \alpha_{1c}^{[F=x_1]}\alpha_{2c}^{[F=x_2]}\cdots\alpha_{kc}^{[F=x_k]}$$

Example:

Categorical class-condi

from nominal data: predicting ***play***

Likelihood of the two classes

For "yes" = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For "no" = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

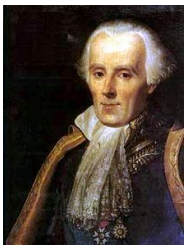
Conversion into a probability by normalization:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

The "Zero-Frequency Problem" (1)

- What if an nominal attribute value doesn't occur with every class value?
  (e.g., no outlook "overcast" for class "no play")
  - Probability will be zero!
  - A posteriori probability will also be zero!
    (no matter how likely the other values are!)
- Remedy for **Categorical likelihood**: add 1 to the count for every attribute value-class combination (*Laplace estimator* or *rule of succession*)
- Result: probabilities will never be zero
  (also stabilizes probability estimates)



Pierre-Simon, marquis de Laplace

The "Zero-Frequency Problem" (2)

(Categorical likelihood continued…)

- In some cases, adding a constant different from 1 might be more appropriate: μ
- Example: attribute *outlook* for class *yes*

Sunny	*Overcast*	*Rainy*
$\frac{2+\frac{\mu}{3}}{9+\mu}$	$\frac{4+\frac{\mu}{3}}{9+\mu}$	$\frac{3+\frac{\mu}{3}}{9+\mu}$

- Weights (p's) don't need to be equal
  (but p's **must sum to 1** !)

$$\frac{2+(\mu \cdot p_1)}{9+\mu} \qquad \frac{4+(\mu \cdot p_2)}{9+\mu} \qquad \frac{3+(\mu \cdot p_3)}{9+\mu}$$

Another Example: Classifying Text

- 20 newsgroup dataset
- Total of 20,000 documents
- Task: assign a new document to one of the 20 newsgroups
- ***Bag-of-words*** model:
  - Total of *M* unique words for all documents
  - $\mathbf{X}_n$ is vector of counts of each of the *M* word types
    - (so $x_{n,m}$ is the number of times word *m* appears in document *n*)

$$p(\mathbf{x}_n|T_n = c, \ldots) = \prod_{m=1}^{M} p(x_{nm}|T_n = c, \ldots)$$

Another Example: Classifying Text

- Note: the *bag-of-words* model ignores word order.

1: The quick brown fox jumps over the lazy dog.

2: Dog quick lazy the jumps fox brown the over.

- **Multinomial** distribution:

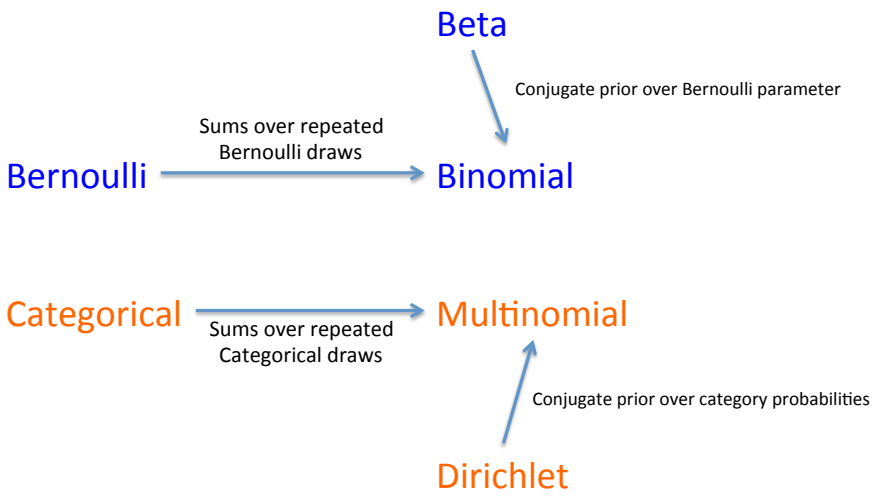
$$P(\mathbf{x}_n|\mathbf{q}) = \left(\frac{s_n!}{\prod_{m=1}^{M} x_{nm}!}\right) \prod_{m=1}^{M} q_m^{x_{nm}}$$

Prior:

$$P(T_{\text{new}} = c|\mathbf{X}, \mathbf{t}) = \frac{1}{C}$$

- Max likelihood estimate of *q*: $q_{cm} = \frac{\sum_{n=1}^{N_c} x_{nm}}{\sum_{m'=1}^{M} \sum_{n=1}^{N_c} x_{nm'}}$
- Prediction:

$$P(T_{\text{new}} = c|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}) = \frac{p(\mathbf{x}_{\text{new}}|T_{\text{new}} = c, \mathbf{X}, \mathbf{t}) P(T_{\text{new}} = c|\mathbf{X}, \mathbf{t})}{\sum_{c'=1}^{C} p(\mathbf{x}_{\text{new}}|T_{\text{new}} = c', \mathbf{X}, \mathbf{t}) P(T_{\text{new}} = c'|\mathbf{X}, \mathbf{t})}$$



The Bayesian Approach to Zero-Frequency problem for Multinomial Likelihood

- The **Dirichlet** Prior:

$$p(\mathbf{q}_c|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{m=1}^{M}\alpha_m\right)}{\prod_{m=1}^{M}\Gamma(\alpha_m)}\prod_{m=1}^{M}q_{cm}^{\alpha_m-1}$$

- MAP estimate of Multinomial Likelihood and Dirichlet prior:

$$q_{cm} = \frac{\alpha - 1 + \sum_{n=1}^{N_c} x_{nm}}{M(\alpha-1) + \sum_{m'=1}^{M}\sum_{n=1}^{N_c} x_{nm'}}$$

- Called "smoothing" because as α is increased, q_{cm} gets closer to $1/M$

Document Classification Results

- Trained on 11,000 documents
- Results of classifying 7,000 held-out documents: