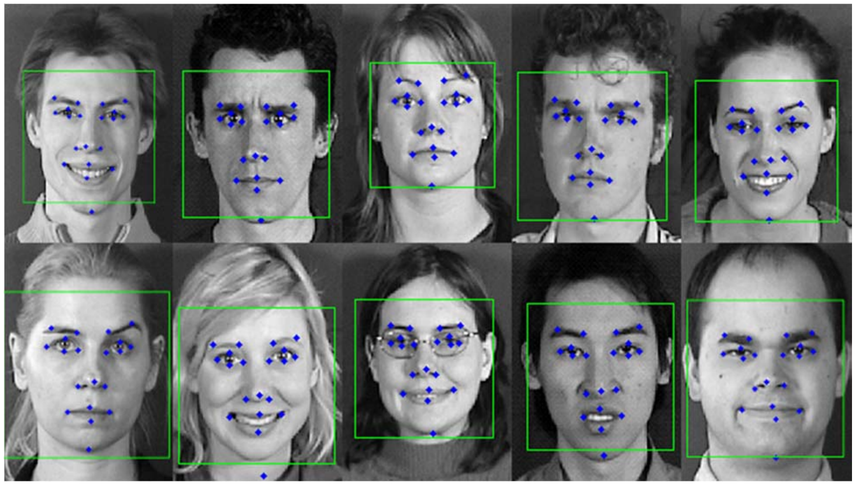
Learning from Data
6. Classification: Naïve Bayes

Spring 2017

Classification

- Supervised (preditive) learning: given examples with labels, predict labels for all unseen examples
  - Classification:
    - label = category,
    - $\mathbf{y} \in \mathcal{Y} = \{1, \ldots, m\}$, m = number of classes
    - $\ell(\mathbf{x}, y, h) = 1(h(\mathbf{x}) \neq y)$, Risk = $P(Y \neq h(\mathbf{X})) = P(\mathrm{Error})$



$\mathbf{x}$ = facial geometry features
y = gender label

Naïve Bayes

- Generative learning
- d features (components) in each feature vector:

$$\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_d)^\top$$

- Naïve Bayes Assumption: all features are conditionally independent given the class label

$$\forall y, \ p(\mathbf{x}|y, \theta) = p(x_1|y, \theta_1) \cdots p(x_d|y, \theta_d) = \prod_{i=1}^{d} p(x_i|y, \theta_i)$$

$$\theta = (\theta_0, \theta_1, \ldots, \theta_d)^\top, \quad \theta_0 = (p(y=1), \ldots, p(y=m))^\top$$

- Resulting MPE decision rule: Naïve Bayes classifier
- Why Naïve? We do not expect all features to be independent even conditional on class label

Naïve Bayes

- Although assumption is false, the resulting classifier works quite well in practice.

- It has only on the order of $O(md)$ parameters to learn ➔ relatively immune to overfitting
  - Example: Suppose that all features are binary, say taking only values +1 or -1, then:
    - In NB model, we need to specify only 1 scalar parameter per feature per class, namely $P(X_i = +1|y), y = 1,\ldots,m$ or md scalar parameters in total
    - But specifying the joint pmf of all d binary features requires $(2^d\text{-}1)$ scalar parameters per class or $m(2^d\text{-}1)$ scalar parameters in total

Naïve Bayes

- Class-conditional likelihood functions:

$$p(x_i|y,\theta_i), i = 1, \ldots, d$$

  - may be either completely parametric or completely non-parametric or some components may be modeled parametrically and others non-parametrically
  - some or all feature components may be discrete, continuous, or mixed
  - Gaussian Naïve Bayes: $\forall i, y,\ p(x_i|y,\theta_i) \sim \mathcal{N}\big(\mu_{iy}, \sigma^2_{iy}\big)$
  - Categorical Naïve Bayes: all features are discrete and take only a finite number of possible values

Naïve Bayes

- Notation summary:

$$y = \text{class label } \in \{1, \ldots, m\}$$

$$j = \text{sample index } \in \{1, \ldots, n\}$$

$$i = \text{feature (component) index } \in \{1, \ldots, d\}$$

$$\theta_i = \text{parameters for class-conditional pdfs/pmfs of feature } i \text{ across all classes}$$

$$\theta_i = \{\theta_{iy}, y = 1, \ldots, m\}, \text{ where}$$

$$\theta_{iy} = \text{parameters for class-conditional pdf/pmf of feature } i \text{ in class } y, \text{ and}$$

$$p(x_i|y, \theta_i) = p(x_i|y, \theta_{iy}) \text{ for all } i, y$$

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \text{ (training set)}$$

feature index $\downarrow$

$$\begin{matrix} i=1 \\ \vdots \\ i \\ \vdots \\ i=d \end{matrix} \qquad \mathbf{x}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{i1} \\ \vdots \\ x_{d1} \end{pmatrix}, \quad \ldots \quad \mathbf{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{dj} \end{pmatrix}, \quad \ldots \quad \mathbf{x}_n = \begin{pmatrix} x_{1n} \\ \vdots \\ x_{in} \\ \vdots \\ x_{dn} \end{pmatrix}$$

$$j = 1, \qquad \ldots \qquad j, \qquad \ldots \qquad j = n$$

sample index $\longrightarrow$

Naïve Bayes

- Training set: $\mathcal{D} = \{(\mathbf{x}_j, y_j), j = 1, \dots, n\}$
- Training set for feature i:
$$\mathcal{D}_i = \{(x_{ij}, y_j), j = 1, \dots, n\}$$
- Training set for feature i and class y:
$$\mathcal{D}_{iy} = \{(x_{ij}, y_j), j: y_j = y\}$$

- Key Result: for each feature i and each class y, the ML estimate of θ_{iy} based on $\mathcal{D}$ is the same as the ML estimate of θ_{iy} based on $\mathcal{D}_{iy}$
- Intuition: features independent in every class $\Rightarrow$ parameters of feature distribution can be estimated independently for each feature in each class

Naïve Bayes

Proof :

$$p(\mathcal{D}|\theta) = \prod_{j=1}^{n} p(\mathbf{x}_j, y_j|\theta)$$

$$= \prod_{j=1}^{n} \left[p(y_j) \prod_{i=1}^{d} p(x_{ij}|y_j, \theta_i) \right]$$

Naïve Bayes assumption:
$p(\mathbf{x}_j|y_j) = \prod_{i=1}^{d} p(x_{ij}|y_j, \theta_i)$

$$= \left(\prod_{j=1}^{n} p(y_j) \right) \prod_{i=1}^{d} \prod_{j=1}^{n} p(x_{ij}|y_j, \theta_i)$$

$$= \left(\prod_{y=1}^{m} p(y)^{n_y} \right) \prod_{i=1}^{d} \prod_{y=1}^{m} \prod_{j:y_j=y} p(x_{ij}|y, \theta_{iy})$$

n_y = # class y samples
$= \sum_{j=1}^{n} 1(y_j = y)$

$$\Rightarrow \ln p(\mathcal{D}|\theta) = \sum_{y=1}^{m} n_y \ln p(y) + \sum_{i=1}^{d} \sum_{y=1}^{m} \sum_{j:y_j=y} \ln p(x_{ij}|y, \theta_{iy})$$

$$\Rightarrow \frac{1}{n} \ln p(\mathcal{D}|\theta) = \sum_{y=1}^{m} \left(\frac{n_y}{n} \right) \ln \theta_{0y} + \sum_{i=1}^{d} \sum_{y=1}^{m} \left(\frac{n_y}{n} \right) \left[\frac{1}{n_y} \sum_{j:y_j=y} \ln p(x_{ij}|y, \theta_{iy}) \right]$$

Naïve Bayes

Thus,

$$\begin{aligned}
\widehat{\theta}_{ML}(\mathcal{D}) &= \arg\max_{\theta} \frac{1}{n} \ln p(\mathcal{D}|\theta), \\
&\theta = (\theta_0, \{\theta_{iy}, i = 1, \ldots, d, y = 1, \ldots, m\}) \\
\Rightarrow \widehat{\theta}_{0,ML}(\mathcal{D}) &= \arg\max_{\theta_0} \left[\sum_{y=1}^{m} \left(\frac{n_y}{n} \right) \ln \theta_{0y} \right], \\
\forall i, y,\ \widehat{\theta}_{iy,ML}(\mathcal{D}) &= \arg\max_{\theta_{iy}} \left[\frac{1}{n_y} \sum_{j: y_j = y} \ln p(x_{ij}|y, \theta_{iy}) \right] \\
&= \widehat{\theta}_{iy,ML}(\mathcal{D}_{iy})
\end{aligned}$$

Categorical Naïve Bayes

- All features are categorical:
  - Without loss of generality (w.l.o.g.): $\forall i,\ x_i \in \{1, \ldots, W\}$
- Notation summary:

$$y = \text{class label } \in \{1, \ldots, m\}$$

$$j = \text{sample index } \in \{1, \ldots, n\}$$

$$i = \text{feature (component) index } \in \{1, \ldots, d\}$$

$$w = \text{feature value } \in \{1, \ldots, W\}$$

- Model:

$$\forall y, j, i, w,\ P(X_{ij} = w | Y_j = y) = \beta_{w,y,i}\text{, where}$$

$$\forall y, i, w,\ \beta_{w,y,i} \geq 0,\ \text{and } \forall y, i,\ \sum_{w=1}^{W} \beta_{w,y,i} = 1$$

- Parameters:

$$\theta = \{\underbrace{p(y)}_{\theta_{0y}}, \underbrace{\beta_{w,y,i}, w = 1, \ldots, W}_{\theta_{iy} = \{\beta_{w,y,i}, w=1,\ldots,W\}}, y = 1, \ldots, m, i = 1, \ldots, d\}$$

- Total number of scalar parameters = $m + mdW$

Categorical Naïve Bayes

- Bayes classifier for 0-1 loss (MPE rule = MAP rule):

$$h_{\mathrm{MPE}}(\mathbf{x}) = \arg\max_{y=1,\dots,m} p(y)p(\mathbf{x}|y,\theta)$$

$$= \arg\max_{y=1,\dots,m} p(y)\prod_{i=1}^{d} \beta_{x_i,y,i}$$

- ML (frequentist) estimate of θ: Let
  - n_y = number of class y training examples
  - $n_{w,y,i}$ = number of class y training examples in which the i-th feature value = w
- then, $\sum_{y=1}^{m} n_y = n$ and for all $i, \sum_{w=1}^{W} n_{w,y,i} = n_y$
- Also, let $n_{w,y} = \sum_{i=1}^{d} n_{w,y,i}$ = number of times w occurs across all d components in class y examples then, $\sum_{w=1}^{W} n_{w,y} = dn_y$

Categorical Naïve Bayes

- ML (frequentist) estimate of θ:

$$\widehat{\theta}_{ML}(\mathcal{D}) = \arg\max_{\theta} \frac{1}{n} \ln p(\mathcal{D}|\theta),$$

$$\theta = (\theta_0, \{\theta_{iy}, i = 1, \ldots, d, y = 1, \ldots, m\})$$

$$\widehat{\theta}_{0,ML}(\mathcal{D}) = \arg\max_{\theta_0} \left[\sum_{y=1}^{m} \left(\frac{n_y}{n} \right) \ln \theta_{0y} \right],$$

$$\begin{aligned}
\forall i, y,\ \widehat{\theta}_{iy,ML}(\mathcal{D}) &= \arg\max_{\theta_{iy}} \left[\frac{1}{n_y} \sum_{j: y_j = y} \ln p(x_{ij}|y, \theta_{iy}) \right] \\
&= \arg\max_{\theta_{iy}} \left[\frac{1}{n_y} \sum_{j: y_j = y} \ln \beta_{x_{ij}, y, i} \right] \\
&= \arg\max_{\theta_{iy}} \left[\sum_{w=1}^{W} \left(\frac{n_{w,y,i}}{n_y} \right) \ln \beta_{w,y,i} \right]
\end{aligned}$$

Categorical Naïve Bayes

- ML (frequentist) estimate of θ solution:

$$\forall y,\ \widehat{\theta}_{0y,ML} = \widehat{p}(y) = \frac{n_y}{n}$$
$$\forall w, y, i,\ \widehat{\beta}_{w,y,i} = \frac{n_{w,y,i}}{n_y}$$

- If for all w,y,i, $\beta_{w,y,i} = \beta_{w,y}$ then,

$$\forall y,\ \widehat{\theta}_{0y,ML} = \widehat{p}(y) = \frac{n_y}{n}$$
$$\forall w, y, i,\ \widehat{\beta}_{w,y,i} = \frac{\sum_{i=1}^{d} n_{w,y,i}}{\sum_{i=1}^{d} n_y}$$
$$= \frac{n_{w,y}}{d \cdot n_y}$$

Categorical Naïve Bayes

- Solution based on following result
- Result: Let $p_1, \ldots, p_L$, and $q_1, \ldots, q_L$, denote two pmfs over L items, i.e., they are non-negative and sum to one. Then,

$$\arg\max_{q_1,\ldots,q_L} \sum_{l=1}^{L} p_l \ln q_l = \{p_l, l = 1, \ldots, L\}$$

- *Proof* : $\forall t > 0, \ln t \leq t - 1$ with equality, if, and only if (iff), $t = 1$. Replacing t with q_l/p_l , multiplying by p_l , and summing over all l, we get

$$\sum_{l=1}^{L} p_l \ln\left(\frac{q_l}{p_l}\right) \leq \sum_{l=1}^{L} p_l \left(\frac{q_l}{p_l} - 1\right) = \sum_{l=1}^{L} \left(q_l - p_l\right) = 1 - 1 = 0.$$

Alternatively, this follows from the fact that the KL-divergence $D(p||q)$ is non-negative and is zero if, and only if, the two pmfs p, q are identical

Overfitting problem

- If $W >> n$, it is quite likely that there is a value w_0 which never occurs in the training set, but occurs in a test example $\mathbf{x}_{\text{test}}$
- For such a w_0, $n_{w_0,y,i} = \beta_{w_0,y,i} = 0$ for all y, i, and $p\left(\mathbf{x}_{\text{test}} \middle| y, \hat{\theta}_{\text{ML}}\right) = 0, \forall y$, and the decision reduces to random guessing. This ignores information from values that were seen in both training and test sets
- Solution 1: remove words that were not seen during training and proceed as before. Better than random guessing, but still ignores information in new words
- Solution 2: Regularize estimation of β by incorporating prior beliefs via a pdf $\pi(\beta)$.

Bayesian Naïve Bayes with Dirichlet prior

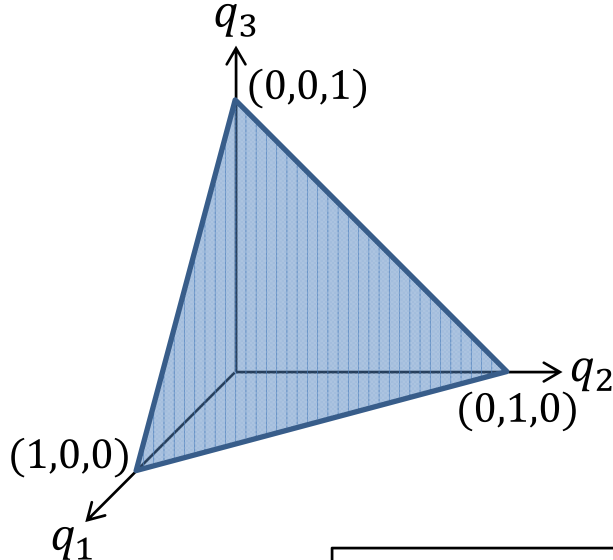
- Dirichlet distribution $\mathrm{Dir}(\mathbf{q}|\boldsymbol{\alpha})$: is a family of continuous multivariate probability distributions parameterized by a W-dimensional vector $\boldsymbol{\alpha}$ of positive reals called the concentration parameters.
- It is a pdf over all pmfs over W values, i.e., a pdf over the (W-1)-dimensional probability simplex:

$$\mathcal{S}_W := \left\{ \mathbf{q} : 0 \leq q_i \leq 1, i = 1, \ldots, W, \sum_{i=1}^{W} q_i = 1 \right\}$$

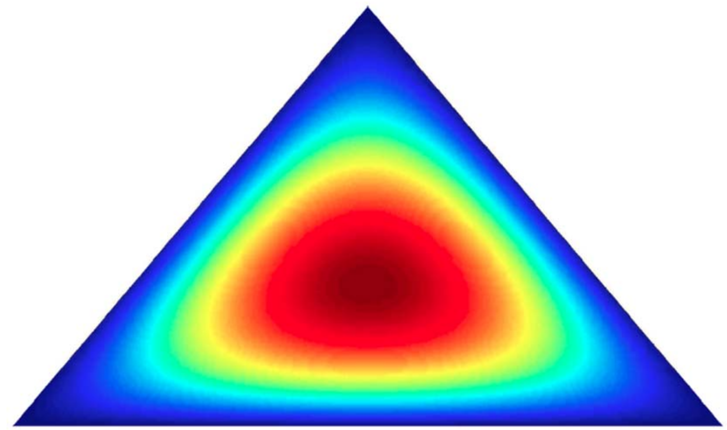
$$\mathrm{Dir}(\mathbf{q}|\boldsymbol{\alpha}) := \frac{1(\mathbf{q} \in \mathcal{S}_W)}{Z(\boldsymbol{\alpha})} \prod_{i=1}^{W} q_i^{\alpha_i - 1}$$

$$Z(\boldsymbol{\alpha}) = \text{normalization constant} \; = \frac{1}{\Gamma\left(\sum_{i=1}^{W} \alpha_i\right)} \prod_{i=1}^{W} \Gamma(\alpha_i)$$

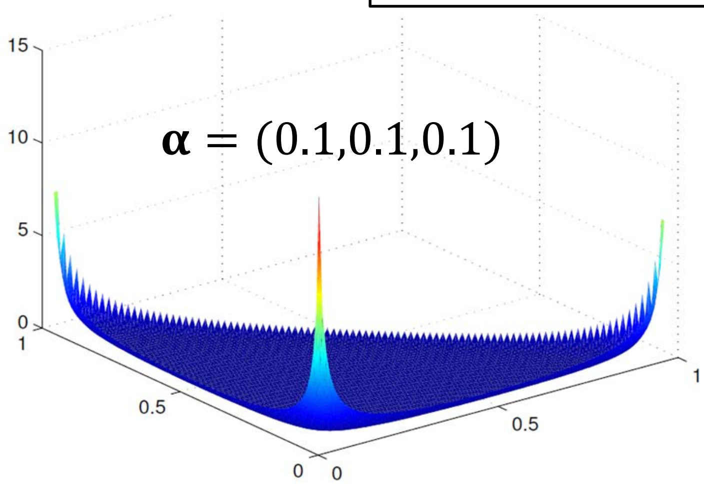
Dirichlet distribution



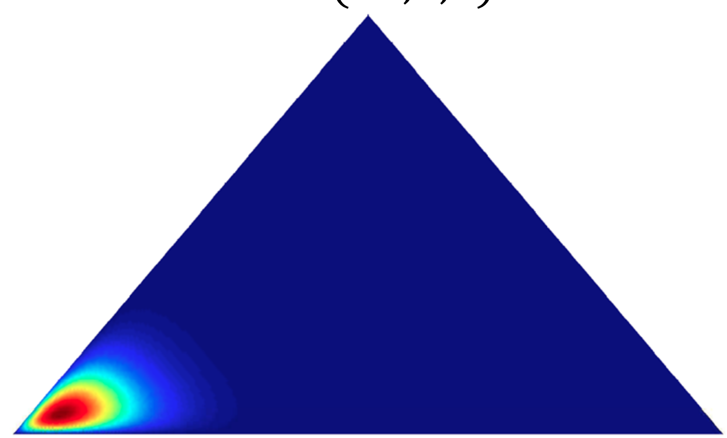
$\boldsymbol{\alpha} = (2,2,2)$



Red (hotter) ➔ larger probability
Blue (cooler)➔ lower probability



$\boldsymbol{\alpha} = (20,2,2)$



Dirichlet distribution

- $\alpha_0 \coloneqq \sum_{i=1}^{W} \alpha_i$ controls how peaked the distribution is (larger ⇒ more peaked)

- Dir(1,1,1) is uniform over the probability simplex

- Dir(2,2,2) is a broad distribution centered at (1/3,1/3,1/3)

- Dir(20,20,20) is a narrow distribution centered at (1/3,1/3,1/3)

- If $\alpha_i < 0$ for all i, we get “spikes” at the corners of the probability simplex

Bayesian Naïve Bayes with Dirichlet prior

- Dirichlet Prior for β:

$$\pi(\beta) = \prod_{y=1}^{m} \prod_{i=1}^{d} \frac{1}{Z(\boldsymbol{\alpha}_{yi})} \prod_{w=1}^{W} \left(\beta_{w,y,i}\right)^{\alpha_{w,y,i}-1}$$

- MAP (Bayesian) estimate of θ:

$$\begin{aligned}
\widehat{\theta}_{\mathrm{MAP}} &= \arg\max_{\theta} p(\theta|\mathcal{D}) \\
&= \arg\max_{\theta} p(\mathcal{D}|\theta)\pi(\beta) \\
&= \arg\max_{\theta} \left[\prod_{y=1}^{m}(p(y))^{n_y}\right] \cdot \left[\prod_{w=1}^{W}\prod_{y=1}^{m}\prod_{i=1}^{d}(\beta_{w,y,i})^{n_{w,y,i}+\alpha_{w,y,i}-1}\right]
\end{aligned}$$

Bayesian Naïve Bayes with Dirichlet prior

- MAP (Bayesian) estimate of θ solution:

$$\forall y,\ \widehat{p}(y) = \frac{n_y}{n}$$
$$\forall w, y, i,\ \widehat{\beta}_{w,y,i} = \frac{n_{w,y,i} + \alpha_{w,y,i} - 1}{n_y + \sum_{w=1}^{W} \left(\alpha_{w,y,i} - 1\right)}$$

- If for all w,y,i, $\beta_{w,y,i} = \beta_{w,y}$ then taking $\alpha_{w,y,i} = \alpha_{w,y}$ for all w,y,i, we get

$$\forall y,\ \widehat{p}(y) = \frac{n_y}{n}$$
$$\forall w, y, i,\ \widehat{\beta}_{w,y,i} = \frac{n_{w,y} + \alpha_{w,y} - 1}{dn_y + \sum_{w=1}^{W} \left(\alpha_{w,y} - 1\right)}$$

Remarks

- alphas can be interpreted as “prior” counts and the MAP solution as updating these prior counts with empirical counts from the likelihood
- If all alphas are equal to 2, then the prior counts are equal to one. This is referred to as add-one smoothing of the ML estimate or Laplace’s rule of succession.
- Can also incorporate a separate Dirichlet prior for $p(y)$ in a similar way
- The ML estimates of θ are asymptotically consistent