Recitation 2
Naive Bayes

Why Bayes Rule?

- Definitions:

  X: Variables we will observe at test time

  Y: Variable we want to predict

- What we want to know:

  $$P(Y \mid X)$$

- What we can compute:

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

There's a catch...

- $P(X \mid Y)$ stands for $P(X_1, \ldots, X_N | Y)$
- If we don't put anything in our model, we don't get anything out.

One Solution (of many)

- Assume:

$$P(X_1, \ldots, X_N | Y) = \prod_{i=1}^{N} P(X_i | Y)$$

- This is a very restrictive assumption!

Inference

- Compute:

$$\underset{Y}{argmax}\, P(Y|X) \propto P(Y) \prod_{i=1}^{N} P(X_i|Y)$$

- For documents, this product may be over a thousand words...

The log-math trick

- Since log is an increasing function,

$$\underset{Y}{argmax}\, P(Y)\prod_{i=1}^{N} P(X_i|Y) =$$

$$\underset{Y}{argmax}\, \log P(Y) + \sum_{i=1}^{N} \log P(X_i|Y)$$

Learning: Parameters are Variables

- Given data, infer $P(X \mid Y), P(Y)$
- Let: θ *index* $\{P(X|Y), P(Y)\}$
- Discrete case:

$$\theta = \{\theta_{ijk} = P(X_i = x_j | Y = y_k), \pi_k = P(Y = y_k)\}$$

- Gaussian case:

$$\theta = \left\{ \begin{matrix} \mu_{ik} = mean(X_i | Y = y_k), \\ \sigma_{ik} = var(X_i | Y = y_k), \pi_k \end{matrix} \right\}$$

- Then $P(\theta | Data) = \dfrac{P(Data|\theta) P(\theta)}{P(Data)}$

Conjugate Priors Example

- Data Distribution:

$$P(X=0|Y=0;\theta)=\theta_0; P(X=1|Y=0;\theta)=1-\theta_0$$

- Say we have N samples where Y=0, of which N_0 cases had X=0 and N_1 had X=1

Conjugate Priors Example

- Beta distribution: $P(\theta_0;\alpha,\beta)=\frac{\theta_0^{\alpha-1}(1-\theta_0)^{\beta-1}}{B(\alpha,\beta)}$

- Bayes rule: $P(\theta_0|Data;\alpha,\beta)=\frac{P(Data|\theta_0)P(\theta_0;\alpha,\beta)}{P(Data)}$

$$=\theta_0^{N0}(1-\theta_0)^{N1}\frac{\theta_0^{\alpha-1}(1-\theta_0)^{\beta-1}}{B(\alpha,\beta)}\frac{1}{C}$$

$$=\frac{\theta_0^{\alpha+N0-1}(1-\theta_0)^{\beta+N1-1}}{B(\alpha+N_0,\beta+N_1)}$$

Conjugate Priors: Another Example

- Data Distribution:

$$P(X=x|Y=0;\mu,\sigma)\propto\exp\left(\frac{-(x-\mu_0)^2}{2\sigma_0^2}\right)$$

- We have one sample where Y=0; X=$\hat{x}$
- Prior over μ_0:

$$P(\mu_0|\tilde{\mu},\tilde{\sigma})\propto\exp\left(\frac{-(\mu_0-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right)$$

Conjugate Priors: Another Example

- Conjugate for μ_0 under known σ_0 is also Gaussian:

$$P(\mu_0 | Data; \tilde{\mu}, \tilde{\sigma}, \sigma_0) \propto P(Data | \mu_0, \sigma_0) P(\mu_0 | \tilde{\mu}, \tilde{\sigma})$$

$$\propto \exp\left(\frac{-(\hat{x}-\mu_0)^2}{2\sigma_0^2}\right)\exp\left(\frac{-(\mu_0-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right)$$

- Complete the square: $\propto \exp\left(\frac{-\left(\mu_0 - \frac{\hat{x}\tilde{\sigma}^2+\tilde{\mu}\sigma_0^2}{\tilde{\sigma}^2+\sigma_0^2}\right)^2}{2(\sigma_0^{-2}+\tilde{\sigma}^{-2})^{-1}}\right)$

MLE's for Gaussians

- Say we have $\{\hat{x}_1, ..., \hat{x}_n\}$ samples of X with the label Y=0.
- MLE for mean: $\hat{\mu}_0 = \frac{1}{n}\sum_i \hat{x}_i$
- MLE for variance: $\hat{\sigma}_0{}^2 = \frac{1}{n}\sum_i (\hat{x}_i - \hat{\mu}_0)^2$

Special Case: Conditionally Dependent Variables

- Want to estimate the bias of a coin
  - Our evidence is photographs of the same coin flip.

- Want to classify the following posting as rec.sport.hockey or talk.politics.misc

  Normally I would keep my postings to rec.sport.hockey, but today I'm here to announce that Sergei Gonchar, who was defenseman for the Pittsburgh Penguins when they won the NHL's Stanley Cup, has decided to run for Governor of Pennsylvania.

Remember the Mushroom Data...

- Which do you think would perform better (One reason for each):
  - Decision Trees
  - Naive Bayes