Jayant Krishnamurthy

10-701 Recitation (9/23/2010)

To cover:

- MLE vs. MAP
  - NB example
  - Priors: Beta (β_H, β_T) vs. Uniform

    add-α smoothing
  - Relation between prior and # of data points
- NB decision boundary
  - Linear vs. quadratic
  - connection to Bayes optimal classifier
- Multinomial & Dirichlet
  - Connection to bernoulli and beta
- LR
  - connection to naive bayes; be careful about conditions.

MLE vs. MAP

MLE and MAP are ways of estimating the parameters of a model

MLE - maximum likelihood estimate (frequentist)

MAP - maximum a posteriori estimate (Bayesian)

Example: Flipping a coin:

$X^{(1)}, \ldots, X^{(n)} \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$ $\quad$ θ: probability coin comes up heads. Heads $\Rightarrow X = 1$

We are given a sample of n flips, $X^1, \ldots, X^n$.

How do we guess the value of θ?

MLE: choose $\hat{\theta}_{MLE}$ to maximize the likelihood of $X^1, \ldots X^n$.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(X^1, \ldots, X^n \mid \theta) \leftarrow \text{data likelihood}$$

$$\hat{\theta}_{MLE} = \arg\max \prod_{i=1}^{n} \theta^{X^i} (1-\theta)^{(1-X^i)} \quad \leftarrow \text{Maximize using Calculus.}$$

In class, we saw $\hat{\theta}_{MLE} = \dfrac{\sum_{i=1}^{n} X^i}{n} = \dfrac{\#\text{ of heads}}{\#\text{ of flips}}$

A difference between frequentists and Bayesians is that Bayesians treat parameters (like θ) as random variables. →

MAP: Use Bayes' Rule to define a distribution over parameters θ:

$$\underbrace{p(\theta \mid X^1, \ldots, X^n)}_{\text{posterior distribution}} = \frac{\overbrace{p(X^1, \ldots, X^n \mid \theta)}^{\text{data likelihood}} \; \overbrace{p(\theta)}^{\text{prior}}}{p(X^1, \ldots, X^n)}$$

Note that $p(\theta \mid X^1, \ldots, X^n)$ is a <u>distribution</u> over possible parameters θ. MAP estimate chooses the θ which maximizes $p(\theta \mid X^1, \ldots, X^n)$:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid X^1, \ldots, X^n)$$

$$= \arg\max_{\theta} p(X^1, \ldots, X^n \mid \theta)\, p(\theta)$$

Issue: we must choose the prior $p(\theta)$
Usually we choose <u>conjugate priors</u> which make MAP estimation easy. Conjugate of Bernoulli is Beta

$$\text{Beta}(\beta_H, \beta_T) = \frac{\theta^{(\beta_H - 1)} (1-\theta)^{(\beta_T - 1)}}{B(\beta_H, \beta_T)}$$

So: $$\hat{\theta}_{MAP} = \arg\max_{\theta} \; \theta^{\beta_H - 1 + \sum_i x_i} (1-\theta)^{\beta_T - 1 + \sum_i (1 - x_i)}$$

And we saw that

$$\hat{\theta}_{MAP} = \frac{(\beta_H - 1) + \sum_i^n x_i}{n + \beta_H + \beta_T - 2}$$ $\Big\}$ β_H and β_T act like "virtual" flips!

Bayesian Inference in Pictures:

In this picture, the dark line is the final posterior distribution. The light lines show what the posterior looks like as we observe more data points. Note that these distributions converge to the final posterior.

A common choice for β_H, β_T is to set $\beta_H = \beta_T = \alpha$, in which case:

$$\hat{\theta}_{MAP} = \frac{(\alpha - 1) + \sum_{i=1}^n x^i}{n + 2(\alpha - 1)}$$ $\Big\}$ This looks like something on your problem set...
(Hint: set $\alpha = 2$)

Multinomial & Dirichlet Distributions

The multinomial distribution is like the Bernoulli distribution for more than 2 values.

Bernoulli - flipping a coin (2 outcomes)
Multinomial - rolling a die (6 outcomes)

Let's say we're rolling an unfair die. How many parameters do we need to describe the distribution? Answer: 5.
In general: n values means $n-1$ parameters.

θ is our parameter vector for the multinomial → θ_i is the probability of observing the ith outcome. (E.g., θ_1 = probability of rolling 1 on the die)

Constraint: $\sum_{i=1}^{n} \theta_i = 1$

Conjugate of multinomial distribution is the dirichlet distribution
(Like conjugate of Bernoulli is Beta)

Note: α is a vector → $\text{Dirichlet}(\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$

As in Beta case, the α_i values act like virtual observations!

$$\hat{\theta}_{MLE} = \frac{\text{Count}(X=i)}{(\text{Total \# of observations})}$$

$$\hat{\theta}_{MAP} = \frac{(\alpha_i - 1) + \text{Count}(X=i)}{(\text{total \# of observations}) + \sum_{i=1}^{n}(\alpha_i - 1)}$$

Naïve Bayes Classifier

Let $X = (X_1, \ldots, X_d)^T$ be a feature vector, $Y \in \{0, 1\}$ be a label.

Recall: classification means predicting Y from X, given training data
$D = \{(X^1, Y^1), (X^2, Y^2), \ldots, (X^n, Y^n)\}$

Naïve Bayes Assumption: Probability of each feature X_i is conditionally independent given Y. Therefore, we can factor distribution $P(Y, X)$:

$$P(X, Y) = \underbrace{P(Y)}_{\text{class prior}} \prod_{i=1}^{d} \underbrace{P(X_i \mid Y)}_{\text{class conditional distributions}}$$

Class-conditional distributions $P(X_i \mid Y)$ may have many different forms. E.g.,

- X_i continuous – Gaussian
- X_i boolean – Bernoulli
- X_i categorical – multinomial

(MLE) Parameter Estimation: (Let $P(Y=1) = \pi$)

$$\hat{\pi} = \frac{1}{n} \sum_{j=1}^{n} Y^j$$

Estimate $P(X_i \mid Y = y)$ based on the distribution.

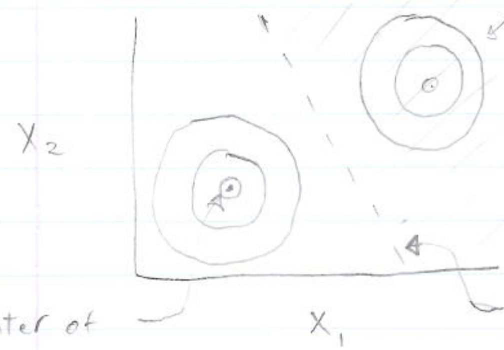
Example: $P(X_i \mid Y)$ is Gaussian with unit variance ($\sigma_i^2 = 1$)

$$P(X_i \mid Y = y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(X_i - \mu_{iy})^2}{2}\right\}$$

$$\hat{\mu}_{iy} = \frac{\sum_{j=1}^{n} \mathbb{1}(Y^j = y)\, X_i^j}{\sum_{j=1}^{n} \mathbb{1}(Y^j = y)}$$

<u>Decision Boundaries</u>

Consider the GNB classifier from the previous example. Assume we have two features X_1, X_2. The classifier will estimate one spherical Gaussian per class. Consider the level sets of these Gaussians:



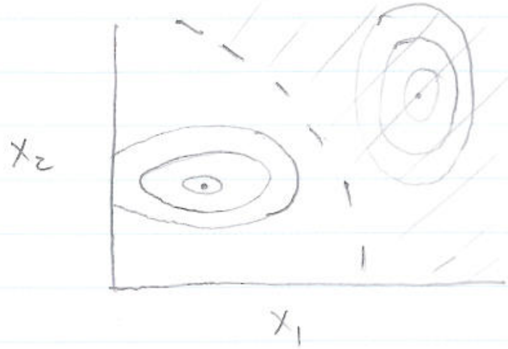
$P(x_1, x_2 | y=1)$

Shaded Region is where the predicted class is 1 (i.e., any instance which falls in the shaded area will have predicted label 1)

Dashed line is the <u>decision boundary</u>, which divides the instance space based on the predicted class.

center of $P(x_1, x_2 | y=0)$

As another example, let's say we now estimate a standard deviation $\hat{\sigma}_{iy}$ for each class conditional distribution:



Notice that the level sets are now axis-aligned ellipses because the $\hat{\sigma}_{iy}$'s scale the dimensions differently.

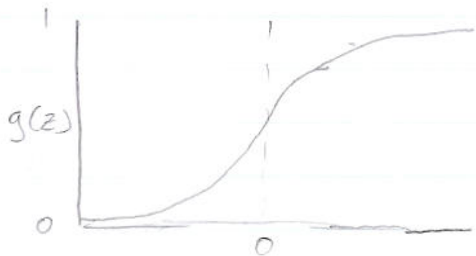
In our first example, our decision boundary is linear, while in the second example it is a curve (actually a quadratic curve). The shape of a classifier's decision boundary is an important characteristic of the classifier.

<u>Logistic Regression (LR):</u>

LR models $P(Y|X)$ instead of $P(X,Y)$.

$$P(Y=1|X) = \frac{\exp\{w_0 + \sum_{i=1}^{n} w_i X_i\}}{1 + \exp\{w_0 + \sum_{i=1}^{n} w_i X_i\}}$$

N.B. $g(z) = \frac{\exp\{z\}}{1+\exp\{z\}}$ is the <u>logistic</u> function.



Logistic function has a value between 0 and 1.
$g(z) \to 0$ as $z \to -\infty$
$g(z) \to 1$ as $z \to \infty$
$g(0) = .5$

LR's decision boundary is linear:

$$P(Y=1|X) = .5 = \frac{\exp\{w_0 + \sum_{i=1}^{n} w_i X_i\}}{1 + \exp\{w_0 + \sum_{i=1}^{n} w_i X_i\}}$$

$$\Rightarrow 0 = w_0 + \sum_{i=1}^{n} w_i X_i.$$

linear equation for the decision boundary.

Where does the LR form come from?
Consider our Gaussian Naive Bayes classifier with unit variance for each class:

$$P(Y=0|X) = \frac{P(Y=0)\,P(X|Y=0)}{P(Y=0)\,P(X|Y=0) + P(Y=1)\,P(X|Y=1)}$$

$$= \frac{1}{1 + \frac{P(Y=1)\,P(X|Y=1)}{P(Y=0)\,P(X|Y=0)}}$$

Consider the second term in the denominator:

$$\frac{P(Y=1)\ P(X \mid Y=1)}{P(Y=0)\ P(X \mid Y=0)} = \exp\left\{ \ln\left(\frac{P(Y=1)}{P(Y=0)} \frac{P(X \mid Y=1)}{P(X \mid Y=0)} \right) \right\}$$

Our NB assumption says $P(X \mid Y) = \prod_{i=1}^{n} P(X_i \mid Y)$
so:

$$\frac{P(Y=1)\ P(X \mid Y=1)}{P(Y=0)\ P(X \mid Y=0)} = \exp\left\{ \ln\left(\frac{P(Y=1)}{P(Y=0)} \right) + \sum_{i=1}^{n} \ln\left(\frac{P(X_i \mid Y=1)}{P(X_i \mid Y=0)} \right) \right\}$$

Consider the terms in the sum:

$$\ln\left(\frac{P(X_i \mid Y=1)}{P(X_i \mid Y=0)} \right) = \ln\left(\frac{\frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} (X_i - \hat{\mu}_{i1})^2 \right\}}{\frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} (X_i - \hat{\mu}_{i0})^2 \right\}} \right)$$

$$= \ln \exp\left\{ \frac{1}{2}(X_i - \hat{\mu}_{i0}) - \frac{1}{2}(X_i - \hat{\mu}_{i1})^2 \right\}$$

$$= \frac{1}{2}\left((X_i^2 - 2X_i\hat{\mu}_{i0} + \hat{\mu}_{i0}^2) - (X_i^2 - 2X_i\hat{\mu}_{i1} + \hat{\mu}_{i1}^2) \right)$$

$$= (\hat{\mu}_{i1} - \hat{\mu}_{i0}) X_i + \frac{\hat{\mu}_{i0}^2 - \hat{\mu}_{i1}^2}{2}$$

Note that the above equation is <u>linear</u> in X_i! So, if we set:

$$w_i = \hat{\mu}_{i1} - \hat{\mu}_{i0}$$

$$w_0 = \ln\left(\frac{P(Y=1)}{P(Y=0} \right) + \sum_{i=1}^{n} \left(\frac{\hat{\mu}_{i0}^2 - \hat{\mu}_{i1}^2}{2} \right)$$

Then:

$$P(Y=0 \mid X) = \frac{1}{1 + \exp\{ w_0 + \sum_{i=1}^{n} w_i X_i \}}$$

WARNING: This derivation does not work for all types of NB classifiers. Example: our earlier GNB classifier with estimated per-class variances. (Recall the decision boundary was a quadratic curve, while LR's decision boundary is always linear!)

Multiclass Logistic Regression

What if we have K labels, $y_1, \ldots, y_K$ instead of just 2 labels?

$$P(Y = y_j | X) = \frac{\exp\{ w_{0j} + \sum_{i=1}^{d} w_{ij} X_i \}}{1 + \sum_{\ell=1}^{K} \exp\{ w_{0\ell} + \sum_{i=1}^{d} w_{i\ell} X_i \}}$$

$$P(Y = y_K | X) = \frac{1}{1 + \sum_{\ell=1}^{K} \exp\{ w_{0\ell} + \sum_{i=1}^{d} w_{i\ell} X_i \}}$$

Decision Rule: Choose y_K with maximum probability.

$$\hat{Y} = \arg\max_{y} P(Y = y | X)$$

$$= \arg\max_{j} \left(0, w_{0j} + \sum_{i=1}^{d} w_{ij} X_i \right)$$

Ex. 3 classes decision boundary

Dashed line is $w_{01} - w_{02} + \sum_{i=1}^{d} (w_{i1} - w_{i2}) X_i = 0$

$w_{02} + \sum_{i=1}^{d} w_{i2} X_i > 0$ (can predict 2

$w_{01} + \sum_{i=1}^{d} w_{i1} X_i > 0$ (can predict 1)