CSC 411: Lecture 08: Generative Models for Classification

Raquel Urtasun & Rich Zemel

University of Toronto

Oct 7, 2015

Today

- Classification – Bayes classifier
- Estimate probability densities from data
- Making decisions: Risk

Generative vs Discriminative

Two approaches to classification:

- Discriminative classifiers estimate parameters of decision boundary/class separator directly from labeled sample
  - learn boundary parameters directly (logistic regression models $p(t_k|\mathbf{x})$)
  - learn mappings from inputs to classes (least-squares, neural nets)

- Generative approach: model the distribution of inputs characteristic of the class (Bayes classifier)
  - Build a model of $p(\mathbf{x}|t_k)$
  - Apply Bayes Rule

Bayes Classifier

- Aim to diagnose whether patient has diabetes: classify into one of two classes (yes C=1; no C=0)
- Run battery of tests
- Given patient's results: $\mathbf{x} = [x_1, x_2, \cdots, x_d]^T$ we want to update class probabilities using Bayes Rule:

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C)p(C)}{p(\mathbf{x})}$$

- More formally

$$\text{posterior} = \frac{\text{Class likelihood} \times \text{prior}}{\text{Evidence}}$$

- How can we compute $p(\mathbf{x})$ for the two class case?

$$p(\mathbf{x}) = p(\mathbf{x}|C=0)p(C=0) + p(\mathbf{x}|C=1)p(C=1)$$

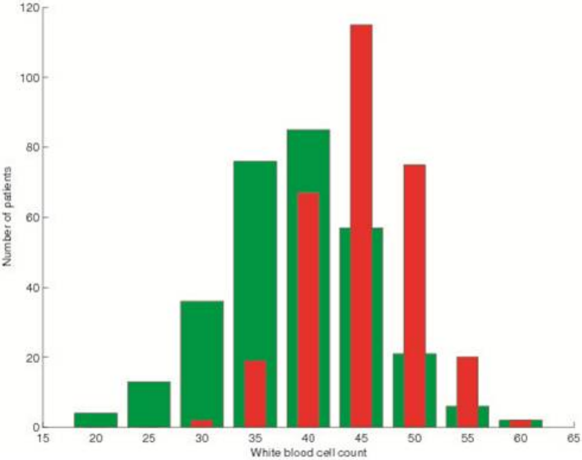
Classification: Diabetes Example

- Start with single input/observation per patient: white blood cell count

$$p(C = 1|x = 50) = \frac{p(x = 50|C = 1)p(C = 1)}{p(x = 50)}$$

- Need class-likelihoods, priors
- Prior: In the absence of any observation, what do I know about the problem?
- What would you use as prior?

Diabetes Data



Question: Which probability distribution makes sense for $p(x|C)$?

MLE for Gaussians

- Let's assume that the class-conditional densities are Gaussian

$$p(x|C) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  with $\mu \in \Re$ and $\sigma^2 \in \Re^+$

- How can I fit a Gaussian distribution to my data?
- Let's try maximum likelihood estimation (MLE)
- We are given a set of training examples $\{x^{(n)}, y^{(n)}\}_{n=1,\cdots N}$ with $y^{(n)} \in \{0,1\}$ and we want to estimate the model parameters $\{\mu, \sigma\}$ for each class
- First divide the training examples into two classes according to $y^{(n)}$, and for each class take all the examples and fit a Gaussian to model $p(x|C)$

MLE for Gaussians II

- We assume that the data points that we have are independent and identically distributed

$$p(x^{(1)}, \cdots, x^{(N)}|C) = \prod_{n=1}^{N} p(x^{(n)}|C) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x^{(n)}-\mu)^2}{2\sigma^2}\right)$$

- Now we want to maximize the likelihood, or minimize its negative (if you think in terms of a loss)

$$\begin{aligned} \ell_{log-loss} &= -\ln p(x^{(1)}, \cdots, x^{(N)}|C) = -\ln\left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x^{(n)}-\mu)^2}{2\sigma^2}\right)\right) \\ &= \sum_{n=1}^{N} \ln(\sqrt{2\pi}\sigma) + \sum_{n=1}^{N} \frac{(x^{(n)}-\mu)^2}{2\sigma^2} = \frac{N}{2}\ln\left(2\pi\sigma^2\right) + \sum_{n=1}^{N} \frac{(x^{(n)}-\mu)^2}{2\sigma^2} \end{aligned}$$

- How would you do we minimize the function?
- Write $\frac{d\ell_{log-loss}}{d\mu}$ and $\frac{d\ell_{log-loss}}{d\sigma^2}$ and equal it to 0 to find the parameters μ and σ^2

Computing the Mean

$$\begin{aligned}\frac{\partial \ell_{log-loss}}{\partial \mu} &= \frac{\partial\left(\frac{N}{2}\ln\left(2\pi\sigma^2\right) + \sum_{n=1}^{N}\frac{(x^{(n)}-\mu)^2}{2\sigma^2}\right)}{\partial \mu} = \frac{d\left(\sum_{n=1}^{N}\frac{(x^{(n)}-\mu)^2}{2\sigma^2}\right)}{d\mu} \\ &= \frac{-\sum_{n=1}^{N}2(x^{(n)}-\mu)}{2\sigma^2} = -\sum_{n=1}^{N}\frac{(x^{(n)}-\mu)}{\sigma^2} = \frac{N\mu - \sum_{n=1}^{N}x^{(n)}}{\sigma^2}\end{aligned}$$

And equating to zero we have

$$\frac{d\ell_{log-loss}}{d\mu} = 0 = \frac{N\mu - \sum_{n=1}^{N}x^{(n)}}{\sigma^2}$$

Thus

$$\boxed{\mu = \frac{1}{N}\sum_{n=1}^{N}x^{(n)}}$$

Computing the Variance

$$\frac{d\ell_{log-loss}}{d\sigma^2} = \frac{d\left(\frac{N}{2}\ln\left(2\pi\sigma^2\right) + \sum_{n=1}^{N}\frac{(x^{(n)}-\mu)^2}{2\sigma^2}\right)}{d\sigma^2}$$

$$= \frac{N}{2}\frac{1}{2\pi\sigma^2}2\pi + \frac{\sum_{n=1}^{N}(x^{(n)}-\mu)^2}{2}\left(\frac{-1}{\sigma^4}\right)$$

$$= \frac{N}{2\sigma^2} - \frac{\sum_{n=1}^{N}(x^{(n)}-\mu)^2}{2\sigma^4}$$

And equating to zero we have

$$\frac{d\ell_{log-loss}}{d\sigma^2} = 0 = \frac{N}{2\sigma^2} - \frac{\sum_{n=1}^{N}(x^{(n)}-\mu)^2}{2\sigma^4} = \frac{N\sigma^2 - \sum_{n=1}^{N}(x^{(n)}-\mu)^2}{2\sigma^4}$$

Thus

$$\boxed{\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)}-\mu)^2}$$

MLE of a Gaussian

- We can compute the parameters in closed form for each class by taking the training points that belong to that class

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}$$

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

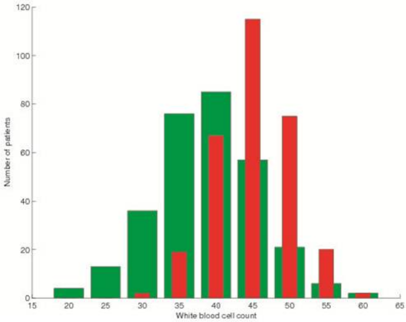
Posterior Probability

- Given a new observation, the estimated class-likelihoods and the prior, we can obtain posterior probability for class $C = 1$

$$
\begin{aligned}
p(C = 1|x) &= \frac{p(x|C = 1)p(C = 1)}{p(x)} \\
&= \frac{p(x|C = 1)p(C = 1)}{p(x|C = 0)p(C = 0) + p(x|C = 1)p(C = 1)}
\end{aligned}
$$

- Lets see an example

Diabetes Example



- Doctor has a prior $p(C = 0) = 0.8$, how?
- Example $x = 50$, $p(x = 50|C = 0) = 0.11$, and $p(x = 50|C = 1) = 0.42$
- How were $p(x = 50|C = 0)$ and $p(x = 50|C = 1)$ computed?
- How can I compute $p(C = 1)$?
- Which class is more likely? Do I have diabetes?

Bayes Classifier

- Use Bayes classifier to classify new patients (unseen test examples)
- Simple Bayes classifier: estimate posterior probability of each class
- What should the decision criterion be?
- The optimal decision is the one that minimizes the expected number of mistakes

Conditional risk of a classifier

$$
\begin{aligned}
R(y|\mathbf{x}) &= \sum_{c=1}^{C} L(y,t)p(t=c|x) \\
&= 0 \cdot p(t=y|x) + 1 \cdot \sum_{c \neq y} p(t=c|x) \\
&= \sum_{c \neq y} p(t=c|x) = 1 - p(t=y|x)
\end{aligned}
$$

- To minimize conditional risk given x, the classifier must decide

$$y = arg \max_c p(t=c|x)$$

- This is the best possible classifier in terms of generalization, i.e. expected misclassification rate on new examples.

Log-odds ratio

- Optimal rule $y = arg\max_c p(t = c|x)$ is equivalent to

$$y = c \quad \Leftrightarrow \quad \frac{p(t = c|x)}{p(t = j|x)} \geq 1 \quad \forall j \neq c$$
$$\Leftrightarrow \quad \log \frac{p(t = c|x)}{p(t = j|x)} \geq 0 \quad \forall j \neq c$$

- For the binary case

$$y = 1 \quad \Leftrightarrow \quad \log \frac{p(t = 1|x)}{p(t = 0|x)} \geq 0$$

- Where have we used this rule before?