02/08/2016

Decision Theory

- Multiple samples with different classes could lie on the same point.
- We want a probabilistic classifier.
- Suppose 10% of the population has cancer; 90% doesn't. We have a probability distribution for calorie intake, $P(X|Y)$:

Calories(X)	< 1200	$1200 - 1600$	> 1600
Cancer (Y=1)	20%	50%	30%
no cancer (Y=-1)	1%	10%	89%

- Recall: $P(X) = P(X|Y = 1)P(Y = 1) + P(X|Y = -1)P(Y = -1)$
- $P(1200 \leq X \leq 1600) = 0.5 \cdot 0.1 + 0.1 \cdot 0.9 = 0.14$
- You meet guy eating $x = 1400$ cals/day. Guess whether he has cancer?
- **Bayes' Theorem**:

$$\begin{aligned} P(A = a|B) &= \frac{P(B|A = a)P(A = a)}{P(B)} \\ P(Y = 1|X = 1400) &= \frac{P(X = 1400|Y = 1)P(Y = 1)}{P(X = 1400)} = \frac{0.05}{0.14} \\ P(Y = -1|X = 1400) &= \frac{P(X = 1400|Y = -1)P(Y = -1)}{P(X = 1400)} \frac{0.09}{0.14} \\ P(Y = 1|X = 1400) &= \frac{5}{14} \approx 36\% \text{ probability guy with 1400 cal/day has cancer} \end{aligned}$$

- A <u>loss function</u> $L(z, y)$ specifies badness if true class is y; classifier prediction is z.

$$L(z, y) = \begin{cases} 1 & \text{if } z = 1, y = -1 \\ 5 & \text{if } z = -1, y = 1 \\ 0 & \text{otherwise} \end{cases}$$

- Definitions:
  - loss function above is <u>asymmetrical</u>
  - The <u>0-1 loss function</u> is 1 for incorrect predictions, 0 for correct.
- Let $r : \mathbb{R}^d \to \pm 1$ be a <u>decision rule</u>, aka <u>classifier</u>: a function that maps a feature vector x to 1 ("in class") or -1 ("not in class").
- The <u>risk</u> for r is the expected loss over all values of x, y:

$$\begin{aligned} R(r) &= E[L(r(X), Y)] \\ &= \sum_x (L(r(x), 1)P(Y = 1|X = x) + L(r(x), -1)P(Y = -1|X = x))P(x) \\ &= \sum_x (L(r(x), 1)\frac{P(X = x|Y = 1)(P(Y = 1)}{P(x)} + L(r(x), 1)\frac{P(X = x|Y = -1)(P(Y = -1)}{P(x)})P(x) \\ &= P(Y = 1) \sum_x L(r(x), 1)P(X = x|Y = 1) + P(Y = -1) \sum_x L(r(x), -1)P(X = x|Y = -1) \end{aligned}$$

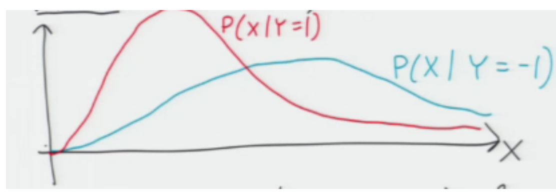
- The Bayes optimal decision rule aka Bayes classifier is the r that minimizes $R(r)$; call it r^*. Assuming $L(z, y) = 0$ for $z = y$:

$$r^*(x) = \begin{cases} 1 & \text{if } L(-1,1)P(Y=1|X=x) > L(1,-1)P(Y=-1|X=x) \\ -1 & \text{otherwise} \end{cases}$$

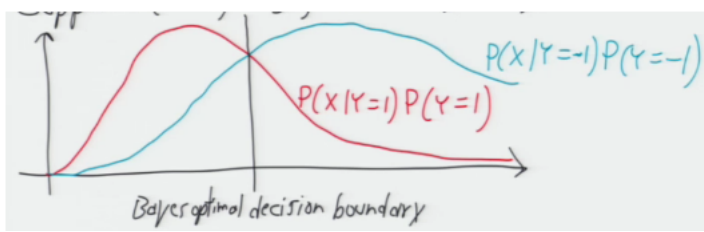
- In cancer example, $r^* = 1$ for all intakes ≤ 1600; $r^* = -1$ for intakes ≥ 1600, then the Bayes risk, aka optimal risk is:

$$R(r^*) = 0.1(5 \cdot 0.3) + 0.9(1 \cdot 0.01 + 1 \cdot 0.1) = 0.249$$

- Suppose X has a continuous probability density function (PDF):
- Review:
  - probability that random variable $X \in [x_1, x_2] = \int_{x_1}^{x_1} P(x)dx$
  - area under whole curve $\int_{-\infty}^{\infty} P(x)dx = 1$
  - expected value of $f(x) : E[f(x)] = \int_{-\infty}^{\infty} f(x)P(x)dx$
  - mean $\mu = E[x] = \int_{-\infty}^{\infty} xP(x)dx$
  - variance $\sigma^2 = E[(X - \mu)^2] = E[X^2] - \mu^2$



  - Suppose $P(Y = 1) = \frac{1}{3}$, $P(Y = -1) = \frac{2}{3}$.



- Define risk as before, replace summations with integrals.

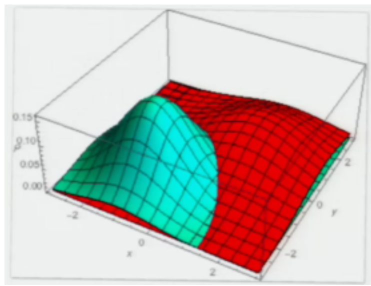
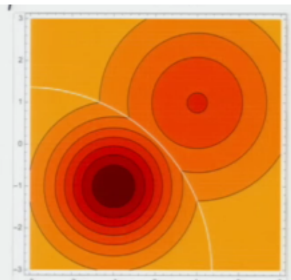
$$\begin{aligned} R(r) &= E[L(r(X), Y)] \\ &= \int_x (L(r(x), 1)P(Y=1|X=x) + L(r(x), -1)P(Y=-1|X=x))P(x)dx \\ &= P(Y=1)\int_x L(r(x), 1)P(X=x|Y=1)dx + P(Y=-1)\int_x L(r(x), -1)P(X=x|Y=-1)dx \end{aligned}$$

- If L is 0-1 loss, $R(r) = P(r(x))$ is wrong

- For Bayes decision rule, Bayes Risk is the area under the minimum of the functions above. Assuming $L(z, y) = 0$ for $x = y$:

$$R(r^*) = \int min_{y\pm 1} L(-y, y) P(X = x|Y = y) P(Y = y) dx$$

- Bayes optimal decision boundary: $\{x : P(Y = 1|X = x) = 0.5\}$.

3 Ways to Build Classifiers

1. Generative models (e.g. LDA)
   - Assume samples come from probability distributions, different for each class.
   - Guess form of distributions.
   - For each class C, fit distribution parameters for class C samples, giving $P(X|Y = C)$.
   - For each C, estimate $P(Y = C)$.
   - Bayes' Theorem gives $P(Y|X)$.
   - If 0-1 loss, pick class C that maximizes $P(Y = C|X = x)$. Equivalently, maximizes $P(X = x|Y = C)P(Y = C)$.
2. Discriminative models (e.g. logistic regression)
   - Model $P(Y|X)$ directly
3. Find decision boundary (e.g. SVM).
   - Model $r(x)$ directly (no posterior).

- Advantages of (1, 2): $P(Y|X)$ tells you probability your guess is wrong.
- Advantage of (1): you can diagnose outliers: $P(x)$ is very small.
- Disadvantages of (1): often hard to estimate distribution accurately; real distributions rarely match standard ones.