6 Decision Theory; Generative and Discriminative Models

DECISION THEORY aka Risk Minimization

[Today I'm going to talk about a style of classifier very different from SVMs. The classifiers we'll cover in the next few weeks are based on probability.]

[One aspect of probabilistic data is that sometimes a point in feature space doesn't have just one class. Suppose your data is adult men and women with just one feature: their height. You want to train a classifier that takes in an adult's height and returns a classification, man or woman. Suppose you are asked to predict the sex of a 5'5" adult. Well, your training set includes some 5'5" women and some 5'5" men. What should you do?]
[In your feature space, you have two training points at the same location with different classes. More generally, the height distributions of men and women overlap. Obviously, in that case, you can't draw a decision boundary that classifies all points with 100% accuracy.]

Multiple sample points with different classes could lie at same point:
we want a probabilistic classifier.

Suppose 10% of population has cancer, 90% doesn't.
Probability distributions for occupation conditioned on cancer, $P(X|Y)$:

job	(X)	miner	farmer	other
cancer	$(Y = 1)$	20%	50%	30%
no cancer	$(Y = -1)$	1%	10%	89%

[caps here mean random variables, not matrices.]

[I made these numbers up. Please don't take them as medical advice.]

Recall: $P(X) = P(X|Y = 1)\,P(Y = 1) + P(X|Y = -1)\,P(Y = -1)$
$P(X = \text{farmer}) = 0.5 \times 0.1 + 0.1 \times 0.9 = 0.14$ [. . . so 14% of random people are farmers]

You meet a farmer. Guess whether he has cancer?

[If you're in a hurry, you might see that 50% of people with cancer are farmers, but only 10% of people with no cancer are farmers, and conclude that a typical farmer probably has cancer. But that would be wrong, because that reasoning fails to take the prior probabilities into account.]

Bayes' Theorem:

$$\begin{aligned} & \downarrow \text{posterior probability} \qquad \downarrow \text{prior prob.} \qquad \downarrow \text{if } X = \text{ farmer} \\ P(Y = 1|X) &= \frac{P(X|Y = 1)P(Y = 1)}{P(X)} = \frac{0.05}{0.14} \\ P(Y = -1|X) &= \frac{P(X|Y = -1)P(Y = -1)}{P(X)} = \frac{0.09}{0.14} \end{aligned}$$

[These two probs always sum to 1.]

$P(\text{cancer} \mid \text{farmer}) = 5/14 \approx 36\%$.

[So we probably shouldn't diagnose cancer.]

[BUT . . . we're assuming that we want to maximize the chance of a correct prediction. But that's not always the right assumption. If you're developing a cheap screening test for cancer, you'd rather have more false positives and fewer false negatives. A false negative might mean somebody misses an early diagnosis and dies of a cancer that could have been treated if caught early. A false positive just means that you spend more money on more accurate tests. When there's an asymmetry between the awfulness of false positives and false negatives, we can quantify that with a loss function.]

A loss function $L(z, y)$ specifies badness if classifier predicts z, true class is y.

E.g., $L(z, y) = \begin{cases} 1 & \text{if } z = 1, y = -1, \quad \text{false positive is bad} \\ 5 & \text{if } z = -1, y = 1, \quad \text{false negative is BAAAAAD} \\ 0 & \text{if } z = y. \quad \text{[loss should \textit{always} be zero for a perfectly correct prediction!]} \end{cases}$

A 36% probability of loss 5 is worse than a 64% prob. of loss 1,
so we recommend further cancer screening.

The loss fn above is asymmetrical.
[A symmetrical loss is the same for false positives and false negatives. For example ...]

The 0-1 loss function is $L(z, y) = \begin{cases} 1 & \text{if } z \neq y, \quad \text{[always 1 for a wrong prediction]} \\ 0 & \text{if } z = y. \quad \text{[always 0 for a correct prediction]} \end{cases}$

[Another application where you want a very asymmetrical loss function, besides medical diagnosis, is spam detection. Putting a good email in the spam folder is much worse than putting spam in your inbox.]

Let $r : \mathbb{R}^d \to \pm 1$ be a decision rule, aka classifier:
a fn that maps a feature vector x to 1 ("in class") or -1 ("not in class").

The risk for r is the expected loss over all values of x, y: [Memorize this definition!]

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

The Bayes decision rule aka Bayes classifier is the fn r^* that minimizes functional $R(r)$.
Assuming $L(1, 1) = L(-1, -1) = 0$,

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

When L is symmetrical, [the big, key principle you should memorize is]
pick the class with the biggest posterior probability.

[But if the loss function is asymmetrical, then you must weight the posteriors with the losses.]
In cancer example, $r^*(\text{miner}) = 1$, $r^*(\text{farmer}) = 1$, and $r^*(\text{other}) = -1$.

The Bayes risk, aka optimal risk, is the risk of the Bayes classifier.
[In our cancer example, the last expression for risk R gives:]

$$R(r^*) = 0.1(5 \times 0.3) + 0.9(1 \times 0.01 + 1 \times 0.1) = 0.249 \qquad \text{No decision rule gives a lower risk.}$$

[It is interesting that, if we really know all these probabilities, we really can construct an ideal probabilistic classifier. But in real applications, we rarely know these probabilities; the best we can do is use statistical methods to estimate them.]

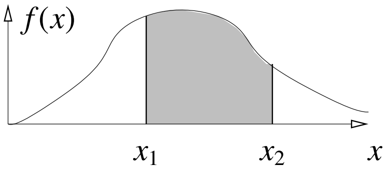
Deriving/using r^* is called risk minimization.

[Did you memorize the two boldfaced lines above yet?]

Continuous Distributions

Suppose X has a continuous probability density fn (PDF).

Review: [Go back to your CS 70 or stats notes if you don't remember this.]



[Draw this by hand. integrate.pdf]

prob. that random variable $X \in [x_1, x_2] = \int_{x_1}^{x_2} f(x)\,\mathrm{d}x$ [shaded area]

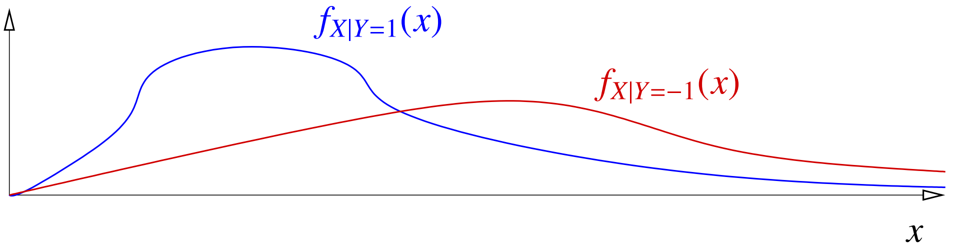
area under whole curve $= 1 = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$

expected value of $g(X)$: $\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x$

mean $\mu = \mathrm{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\,\mathrm{d}x$

variance $\sigma^2 = \mathrm{E}[(X - \mu)^2] = \mathrm{E}[X^2] - \mu^2$

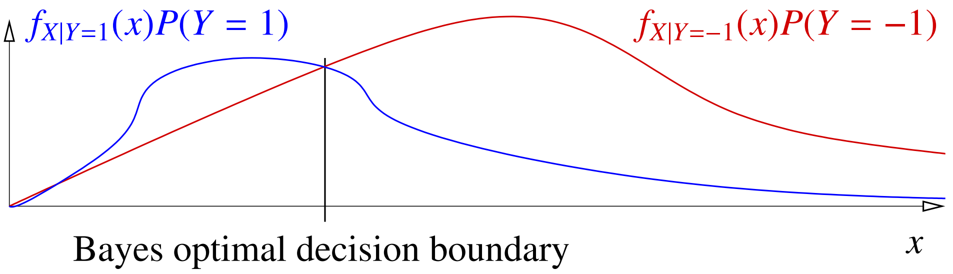
[Perhaps our cancer statistics look like this.]



Draw this figure by hand (cancerconditional.png) [The area under each curve is 1.]

[Let's use the 0-1 loss function. In other words, suppose you want a classifier that maximizes the chance of a correct prediction. The wrong answer would be to look where these two curves cross and make that be the decision boundary. As before, it's wrong because it doesn't take into account the prior probabilities.]

Suppose $P(Y = 1) = 1/3$, $P(Y = -1) = 2/3$, 0-1 loss.



Draw this figure by hand (cancerposterior.png)

[To maximize the chance you'll predict correctly whether somebody has cancer, the Bayes decision rule looks up x on this chart and picks the curve with the highest probability. In this example, that means you pick cancer when x is left of the optimal decision boundary, and no cancer when x is to the right.]

Define risk as before, replacing summations with integrals.

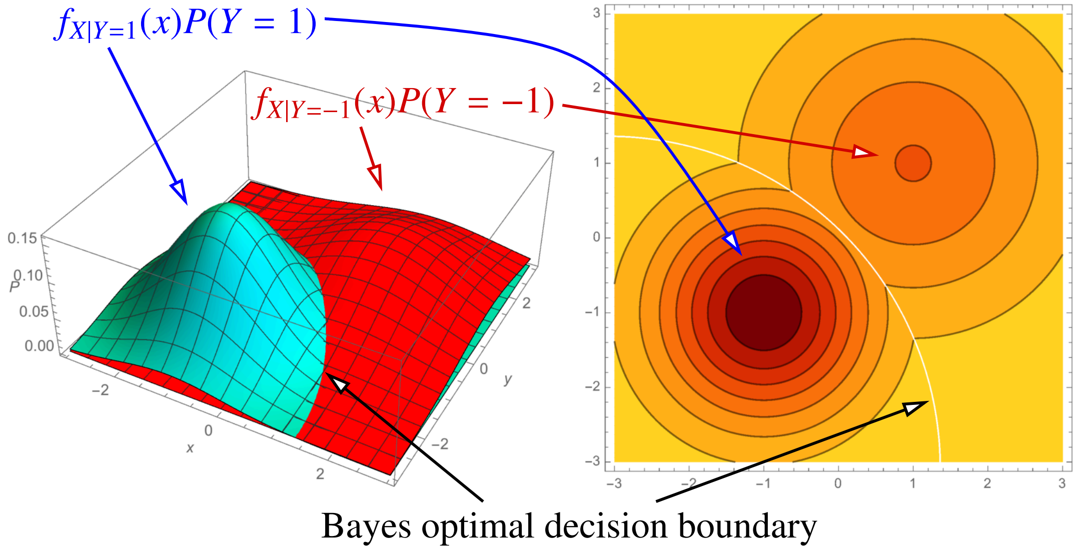
$$\begin{aligned} R(r) &= \mathrm{E}[L(r(X), Y)] \\ &= P(Y=1) \int L(r(x), 1)\, f_{X|Y=1}(x)\, \mathrm{d}x + \\ & P(Y=-1) \int L(r(x), -1)\, f_{X|Y=-1}(x)\, \mathrm{d}x. \end{aligned}$$

For Bayes decision rule, Bayes risk is the area under minimum of functions above. [Shade it.]
Assuming $L(1,1) = L(-1,-1) = 0$,

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

[If you want to use an asymmetrical loss function, just scale the curves vertically in the figure above.]

If L is 0-1 loss, [then the risk has a particularly nice interpretation:]
$R(r) = P(r(x)$ is wrong) [which makes sense, because R is the expected loss.]
and the Bayes optimal decision boundary is $\{x : \underbrace{P(Y=1|X=x)}_{\text{decision fn}} = \underbrace{\;0.5\;}_{\text{isovalue}}\}$



qda3d.pdf, qdacontour.pdf [Two different views of the same 2D Gaussians.]

[Notice that the accuracy of the probabilities is most important near the decision boundary. Far away from the decision boundary, a bit of error in the probabilities probably wouldn't change the classification.]

[You can also have multi-class classifiers, choosing among three or more classes. The Bayesian approach is a particularly convenient way to generate multi-class classifiers, because you can simply choose whichever class has the greatest posterior probability. Then the decision boundary lies wherever two or more classes are tied for the highest probability.]

3 WAYS TO BUILD CLASSIFIERS

(1) Generative models (e.g., LDA) [We'll learn about LDA next lecture.]
- Assume sample points come from probability distributions, different for each class.
- Guess form of distributions
- For each class C, fit distribution parameters to class C points, giving $f_{X|Y=\mathrm{C}}(x)$
- For each C, estimate $P(Y = \mathrm{C})$
- Bayes' Theorem gives $P(Y|X)$
- If 0-1 loss, pick class C that maximizes $P(Y = \mathrm{C}|X = x)$ [posterior probability]
  equivalently, maximizes $f_{X|Y=\mathrm{C}}(x)\, P(Y = \mathrm{C})$

(2) Discriminative models (e.g., logistic regression)
[We'll learn about logistic regression in a few weeks.]
- Model $P(Y|X)$ directly

(3) Find decision boundary (e.g., SVM)
- Model $r(x)$ directly (no posterior)

Advantage of (1 & 2): $P(Y|X)$ tells you probability your guess is wrong
[This is something SVMs don't do.]
Advantage of (1): you can diagnose outliers: $f(x)$ is very small
Disadvantages of (1): often hard to estimate distributions accurately;
real distributions rarely match standard ones.

[What I've written here doesn't actually define the phrases "generative model" or "discriminative model." The proper definitions accord with the way statisticians think about models. A generative model is a full probabilistic model of all variables, whereas a discriminative model provides a model only for the target variables that we want to predict.]

[It's important to remember that we rarely know precisely the value of any of these probabilities. There is usually error in all of these probabilities. In practice, generative models are most popular when you have phenomena that are well approximated by the normal distribution or another "nice" distribution. Generative methods also tend to be more stable than other methods when the number of training points is small or when there are a lot of outliers.]