Announcements

Assignment two due Friday, Feb 10.

This week:

1) I am away Wednesday through Friday
2) Thursday class will be taken by Kyle
3) No Friday office hour

Plan for today:

Push through modeling concepts so Kyle can start on Graphical Models.

Part 02 preview now posted:

Model Selection

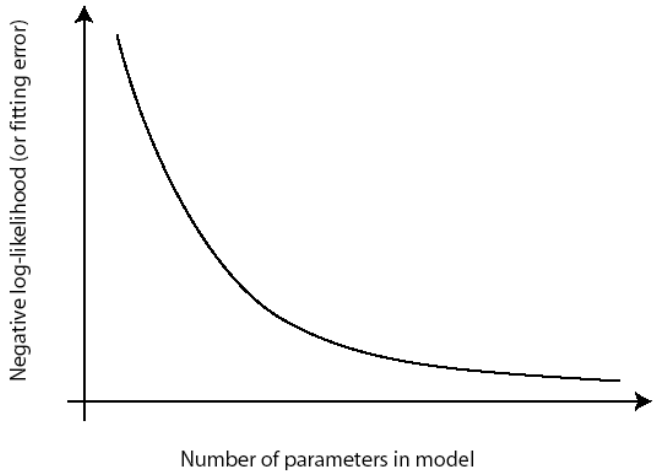
- Model selection refers to choosing among different instances within a model class (1) or different model classes (2).
- Examples:
  - The number of clusters (1)
  - The degree of a polynomial to fit a curve to data (1)
  - Polynomials versus other basis functions such as Fourier (2)

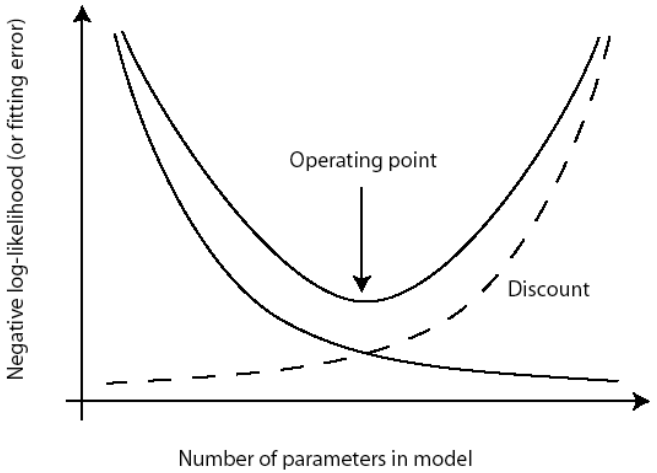
Model Comparison Difficulties

- Prior densities of different models are typically of different dimensionality (leads to expensive integration).
- Good likelihoods help select models, but constructing them is an exacting task.
  - Don't forget about the "negative space"
    - A more complex model (e.g., more objects in a scene) explains more data, but it also proposes more data where there is none.
    - Missing data must be penalized!
- Good priors over different model classes are often not obvious

Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.





Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.
- AIC (An information criterion, Akaike, 74)

  Replace log likelihood, $\log(p(D|\theta))$, with $\log(p(D|\theta)) - M$
  where M is the number of adjustable parameters.

Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.
- BIC (Bayesian information criterion)

  Replace log likelihood, $\log(p(D|\theta))$, with $\log(p(D|\theta)) - \frac{1}{2} M \log(N)$
  where M is the number of adjustable parameters, N is the number of data points. This is the usual approximation. See Bishop, page 216-217 for a more complicated version.

  Often also called minimum discription length (MDL)

A question raised in class revealed that the the dependency on N is confusing. This is resolved by noting that the likelihood typically depends on N (often N is an exponent), but the formula above does not expose this.

Solutions (likelihood function)

- Incorrect complex models may predict lots of data where there is none
- Solution is to model missing data
- Example --- finding asteroids from detections amidst noise
  - Predicting more asteroids explains more data, but we expect to see detections for them most of the time.
  - Good modeling the probability of noise detections and probability of missing detections has a greater affect on the posterior than a prior (necessarily not very strong) on the number of asteroids.

Solutions (integrating parameter uncertainty)

$$p(D|M_i) = \int_{\Omega_i} p(D|\theta)\, p(\theta|M_i)\, d\theta$$ (Model evidence)

and we can evaluate $p(M_i|D)$ by Bayes.

The dimension of the space of θ (Ω_i in the integral) is typically a function of i.

This is argued (Bishop, §3.4) to be a principled way to penalize complex models because complex models spread their probability mass over greater support.

Under additional approximations and assumptions, this becomes BIC (Bishop, §4.4.1).

Solutions (model averaging)

Recall the predictive distributions

$$p(x|X) = \int p(x|\theta)\, p(\theta|X)\, d\theta$$

To mitigate uncertainty of different models

$$p(x|X) = \sum_i p(M_i) \int_{\Omega_i} p(x|\theta_i)\, p(\theta_i|X, M_i)\, d\theta$$

Note the assumption that M_i influences x through θ_i only, so no conditioning on M_i in the first factor in the integral.

Cross-validation

- Standard way to evaluate models
- Exclude a subset of the data while fitting model
- Compute predictions for the held-out subset.
- Evaluate predictions against actual held-out values
  - e.g., distance from truth, or class labels
- If you use k such sets, this is called k-fold cross-validation
- If you leave out 1 data point, it is called leave-one-out.

Cross-validation (2)

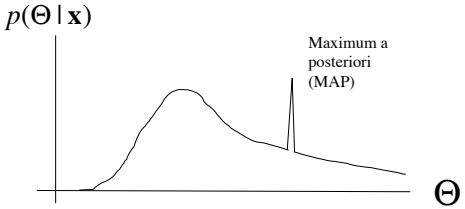
- Cross-validation provides
  - A way to choose models
  - A way to measure performance
  - A way to measure generalization capacity
- Held out data **must be different enough** to test the level of generality that you want
  - Consider degree of validation in a model to predict happiness
    1. How happy are you now given recent data points
    2. How happy are you now given all data points
    3. How happy are you on day X given data for other days
    4. How happy are you based on model of **other** people
    5. How happy are you based on **other** people in other experiments
    6. How happy are you based on modeling people in other cultures

The following slides are skipped for now, so we can do graphical models on Thursday. They are included here in case we do not get back to this.

More on estimation

Skipped in 2012.

- If the goal is to provide the model, then we typically estimate the MAP value for the parameters
- This assumes that the posterior is nicely behaved
- An alternative is to average some or all (MMSE) of the posterior.



Classification

Skipped in 2012.

- Consider that our parameters include a discrete class variable, c.
- Assume no other variables, or that they have been marginalized out.
- Use x for the data.

$$p(c \mid x) \propto p(c)p(x \mid c)$$

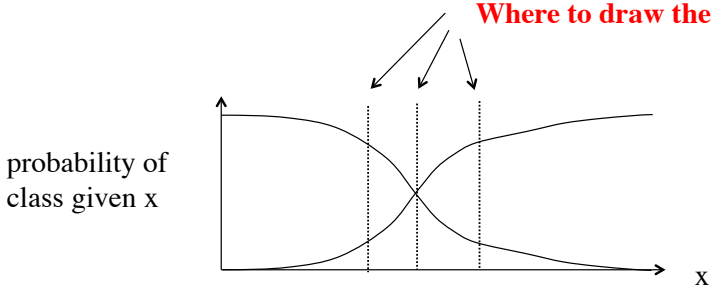
- So, given x, what is the class?

Classification

Skipped in 2012.

Binary case, easy to draw
Two classes, C_1 and C_2.
being in one is the same as **not** being in the other.

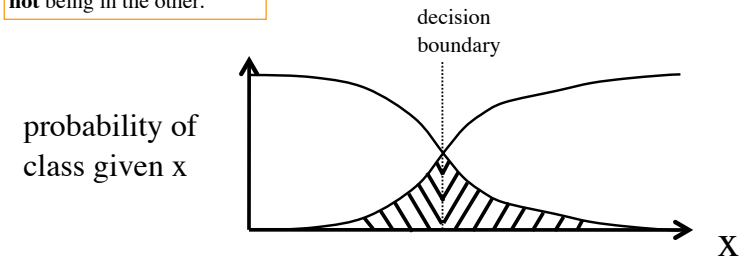
Where to draw the line?



Classification

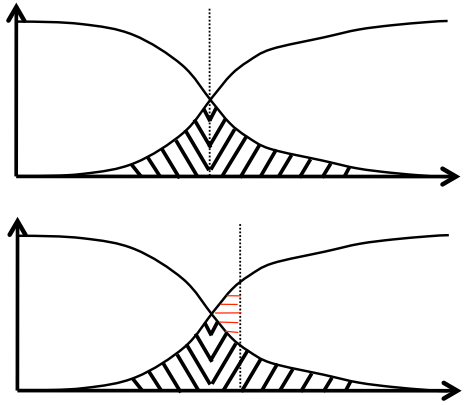
Skipped in 2012.

Binary case, easy to draw
Two classes, C_1 and C_2.
being in one is the same as **not** being in the other.



Area of intersection under curves gives expected value of making a mistake

Skipped in 2012.

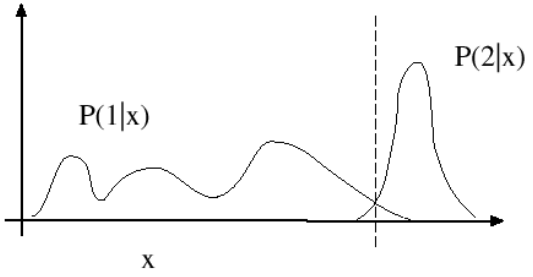


Red shows extra that you get wrong with different boundary

Classification

Skipped in 2012.

Finding a decision boundary is not the same as modeling a conditional density.



Here there are more than two classes, but only two shown. Consider all animals, but you are being force to choose between “dog” and “cat”.

Classification

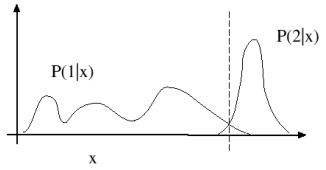
Skipped in 2012.

Finding a decision boundary is not the same as modeling a conditional density.

Working with the boundary might be easier (we don't care about the extra bumps).

But we loose any indication of whether the point is an outlier.

In this course we will not cover in detail finding the boundary (discriminative method).



Bishop §1.5

Decision making

Skipped in 2012.

Classification where the risk (loss) for each class is different.

Example: Risk of a false negative diagnosis is more than that for the risk of false positive diagnosis.

Define a loss function, Lj,k which tells us the loss of classifying a category k, as a category, j.

Example:

	cancer	normal
cancer	0	1000
normal	1	0

Decision making

Skipped in 2012.

Now the classification boundaries for **x** are based on the loss, not just the probability.

You choice of the class, j, for x is the lowest expected loss.

This is found by:

$$\underset{j}{\operatorname{argmin}}\left\{\sum_{k} L_{k,j}\cdot p(C_k|x)\right\}$$

Decision making

Skipped in 2012.

Example to illustrate that the formula is sensible.

Suppose that at a given x*, we have

$p(C_1 \mid x^*)=0.3 \quad p(C_2 \mid x^*)=0.2 \quad p(C_3 \mid x^*)=0.5$

Evaluate the assignment of x* under loss functions

$$L_A=\begin{pmatrix}0&1&1\\1&0&1\\1&1&0\end{pmatrix} \qquad L_B=\begin{pmatrix}0&1&1\\10&0&10\\1&1&0\end{pmatrix}$$

Skipped in 2012.

$p(C_1 \mid x^*) = 0.3 \quad p(C_2 \mid x^*) = 0.2 \quad p(C_3 \mid x^*) = 0.5$

For the first example (loss is misclassifcation rate)

$$L_B = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Note that loss is defined for misclassifying the column item as the row item.

Declaring that x at x* is C_1 has expected loss: (0.3)*0 +(0.2)*1+ (0.5)*1=0.7

Declaring that x at x* is C_2 has expected loss: (0.3)*1 +(0.2)*0+ (0.5)*1=0.8

Declaring that x at x* is C_3 has expected loss: (0.3)*1 +(0.2)*1+ (0.5)*0=0.5

As expected, the minimum loss is for the likeliest class.

Skipped in 2012.

$p(C_1 \mid x^*) = 0.3 \quad p(C_2 \mid x^*) = 0.2 \quad p(C_3 \mid x^*) = 0.5$

For the second example

$$L_B = \begin{pmatrix} 0 & 1 & 1 \\ 10 & 0 & 10 \\ 1 & 1 & 0 \end{pmatrix}$$

Note that loss is defined for misclassifying the column item as the row item.

Declaring that x at x* is C_1 has expected loss: (0.3)*0 +(0.2)*10+ (0.5)*1 = 2.5

Declaring that x at x* is C_2 has expected loss: (0.3)*1 +(0.2)*0+ (0.5)*1 = 0.8

Declaring that x at x* is C_3 has expected loss: (0.3)*1 +(0.2)*10+ (0.5)*0 = 2.3

Now the heavy penalty for missing C_2 leads to C_2 being the best answer.
(Note that C_2 was the worst answer with the previous loss).