Basic concepts
Introduction to Bayesian methods

Lecture 1c

Probabilities

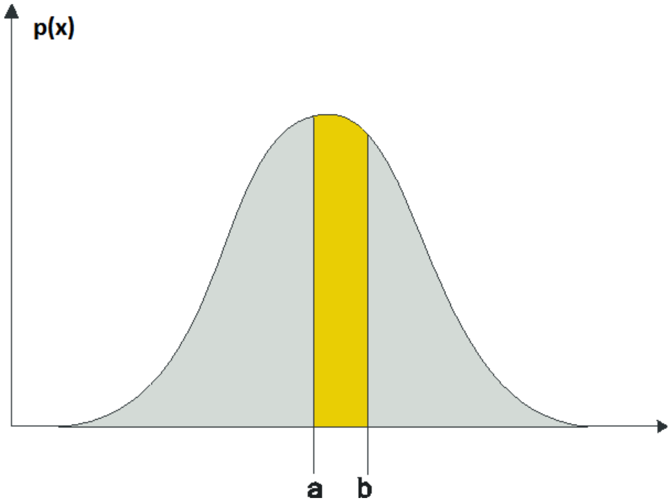
- Frequentist tradition: probabilities derived from counts

Example: Tossing two dices

- X=number on first dice
- Y=number on second dice

- $p(x = X)$ frequency of observing X
- $p(x = X, y = Y)$ frequency of observing X and Y
- $p(x = X|y = Y)$ frequency of observing X given y=Y

Probability density



- $p(x \in [a, b]) = \int_a^b p(x)dx$
- $p(x) \geq 0, \int_{-\infty}^{+\infty} p(x)dx = 1$

Probabilities

- **Laws of probabilities**
  - Sum rule (compute **marginal** probability)

$$p(X) = \sum_Y p(X,Y)$$

$$p(X) = \int p(X,Y)dY$$

  - Product rule

$$p(X,Y) = p(X|Y)p(Y)$$

Combination 1:

$$p(X) = \sum_Y p(X|Y)p(Y)$$

$$p(X) = \int p(X|Y)p(Y)dY$$

Bayes theorem

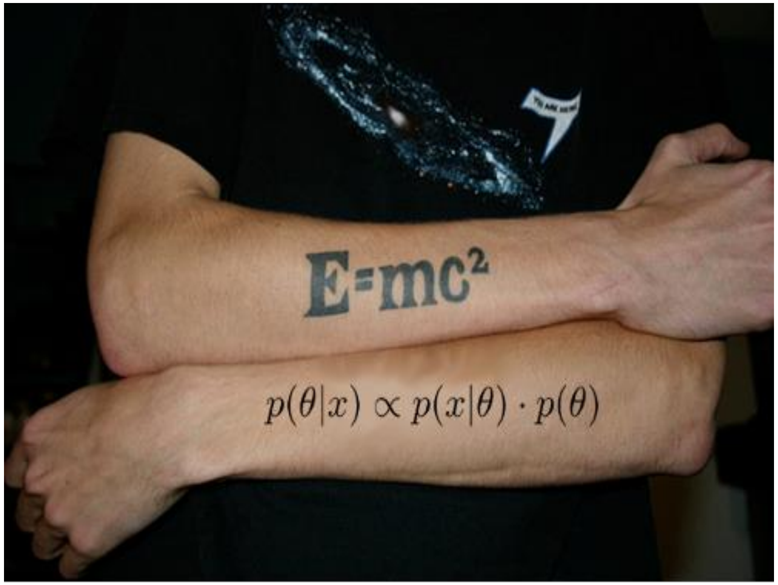
- Combination 2:

Bayes Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(Y|X) \propto p(X|Y)p(Y)$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{\int p(X|Y)p(Y)dY}$$



Bayesian probabilities

- Probability reflects your knowledge (uncertainty) about a phenomenon → **subjective probabilities**
  - **Prior probability** $p(w)$, can be uninformative $p(w) \propto 1$
  - Formulate a model, compute **likelihood** $p(D|w)$
  - **Posterior probability** $p(w|D)$, after observing data
    - $p(w|D) \propto p(D|w)p(w)$

- Model parameters are considered as random variables
  - In real life, do not need to be random, but we model as random

Basic ML ingridients

- Data D: observations
  - Features $X_1, .. X_p$
  - Targets $Y_1, \dots, Y_r$

Case	X_1	X_2	Y
1			
2			
…			

- Model $P(x|\, w_1, \dots w_k)$ or $P(y|x, w_1, \dots w_k)$
  - Example: Linear regression $p(y|x, w) = N(w_0 + w_1 x, \sigma^2)$

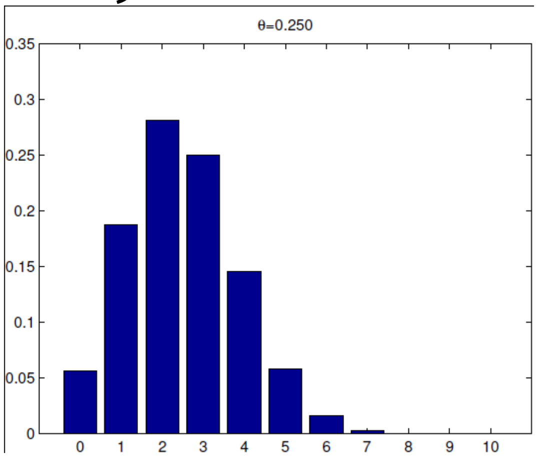
- Learning procedure (data→get parameters $\hat{w}$ or $p(w|D)$)
  - Maximum likelihood, MAP, Bayes rule…

- Predict new data X^{new} by using the fitted model

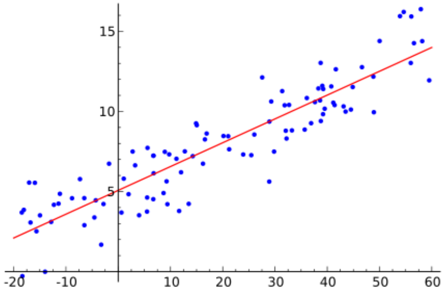
Probabilistic models

- A distribution $p(x|w)$ or $p(y|x,w)$
- Example:
  - $x \sim Bin(n, \theta)$

$$p(x = k|n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$



  - $y \sim N(\alpha_0 + \alpha_1 x, \sigma^2)$



Source: Wikipedia

Learn basic distributions and their properties→PRML, chapter 2!

Fitting a model

- Given dataset D and model $p(\boldsymbol{x}|w)$ or $p(y|\boldsymbol{x},w)$
  - Frequentist approach: which combination of parameter values fits my data best?
  - Bayesian approach: parameters are random variables, all feasible values are acceptable
    - Different parameter values have different probabilities

Fitting a model

- Frequenist principle: Maximum likelihood principle
  - Compute likelihood $p(\boldsymbol{D}|\ w)$

    $p(\boldsymbol{D}|\ w) = \prod_{i=1}^{n} p(X_i|w)$ or $p(\boldsymbol{D}|\ w) = \prod_{i=1}^{n} p(Y_i|X_i, w)$
  - Maximize the likelihood and find the optimal w^* → they are the fitted values

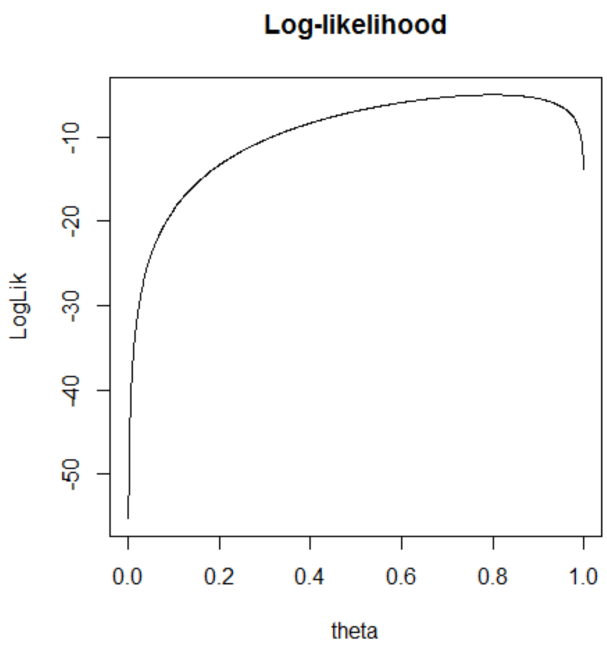
Remarks:

- Likelihood shows how much the chosen parameter value is proper for a specific model and the given data
- Normally **log-likelihood** is used in computations instead
- Other alternatives to ML exist…

Fitting a model

Example: tossing a coin.

$$D = \{0,1,1,0,1,1,1,1,1,1,1,1\},$$
$$p(x = 1|\theta) = \theta, p(x = 0|\theta) = 1 - \theta$$



http://cdn.toonvectors.com/images/35/10267/toonvectors-10267-940.jpg

Fitting a model

- Bayesian principle
  - Compute $p(w|D)$and then decide yourself what to do with this (for ex. MAP, mean, median)
- Use bayes theorem

$$\boxed{p(w|D) = \frac{p(D|w)p(w)}{p(D)} \propto p(D|w)p(w)}$$

- $p(D)$ is **marginal likelihood**
  - $p(D) = \int p(D|w)p(w)dw$ or
  - $p(D) = \sum_i p(D|w_i)p(w_i)$

Example: tossing a coin. Find $p(\theta|D)$, estimate various θ^*

Fitting a model

- How to chose the prior?
  - Expert knowledge about the phenomenon
  - Forcing a model to have a certain structure
    - Example: decision trees: prior prefers smaller trees
  - Conjugacy http://en.wikipedia.org/wiki/Conjugate_prior
    - Distribution of the posterior is the same type as the distribution of the likelihood or prior
- Prior is the most controversial about Bayesian methods, but
  - When $N \rightarrow \infty$, data overwhelms the prior

Prediction

- **Plug-in estimation** (Frequentist and Bayesian)
  - Substitute the estimated w^* into $p(\boldsymbol{x}|w)$ or $p(y|\boldsymbol{x}^{\boldsymbol{new}}, w)$
- **Bayesian model averaging**
  - Posterior predictive distribution:
    - $p(\boldsymbol{x}^{new}|D) = \int p(\boldsymbol{x}^{new}|w)p(w|D)dw$
    - $p(\boldsymbol{y}|D) = \int p(\boldsymbol{y}|w, \boldsymbol{x}^{new})p(w|D, \boldsymbol{x}^{new})dw$

Black swan paradox

- In the coin example, $p(x^{new} = 1) = \frac{k}{n}$ if MLE used
- If we made 3 attempts, no successes→ k=0
- Does this mean $p(x^{new} = 1) = 0$??

- Problem does not appear in Bayesian setting (posterior mean)

Types of supervised models

- **Generative models**: model $p(X|Y,w)$and $p(Y|w)$
  - Example: k-NN classification

$$p(X=x|Y=C_i,K)=\frac{K_i}{N_iV},p(C_i|K)=\frac{N_i}{N}$$

From Bayes Theorem,

$$p(Y=C_i|x,K)\propto\frac{K_i}{K}$$

- **Discriminative models**: model $p(Y|X,w)$, X constant
  - Example: logistic regression
  - $p(Y=1|\boldsymbol{w},\boldsymbol{x})=\frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}}$

Generative vs Discriminative

- Generative can be used to generate new data
- Generative normally easier to fit (check Logistic vs K-NN)
- Generative: each class estimated separately→ do not need to retrain when a new class added
- Discriminative models: can replace X with $\phi(X)$ (preprocessing), method will still work
  - Not generative, distribution will change
- Generative: often make too strong assumptions about $p(X|Y,w)$ → bad performance

Bayesian decision theory

- Machine learning models estimate $p(y|x)$ or $p(y|x, \hat{w})$
- Transform probability into action→ which value to predict?→decision step
  - $p(Y = Spam|x) = 0.83$→do we move the mail to Junk?
  - What is more dangerous: deleting 1 non-spam mail or letting 1 spam mail enter Inbox?
- →**Loss function** or **Loss matrix**

Loss matrix

- Costs of classifying $Y = C_k$ to C_j:
  - Rows: true, columns: predicted

$$L = \left\| L_{ij} \right\|, i = 1, \ldots, n, j = 1, \ldots, n$$

- Example 1: 0/1-loss

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

- Example 2: Spam

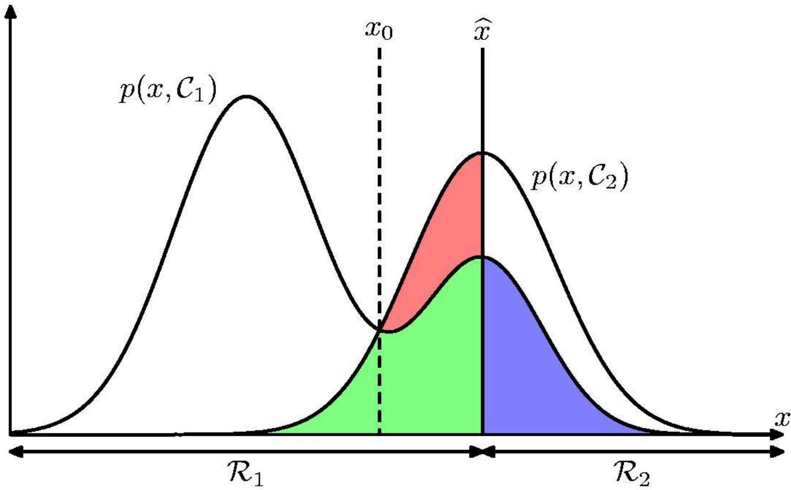
$$L = \begin{pmatrix} 0 & 100 \\ 1 & 0 \end{pmatrix}$$

Loss and decision

- Expected loss minimization
  - R_j : classify to C_j

$$EL = \sum_k \sum_j \int_{R_j} L_{kj} p(\boldsymbol{x}, C_k) d\boldsymbol{x}$$

- **Choose such $\boldsymbol{R_j}$ that $\boldsymbol{EL}$ is minimized**
- Two classes



$$EL = \int_{R_1} L_{21} p(x, C_2) dx + \int_{R_2} L_{12} p(x, C_1)\, dx$$

Loss and decision

- How to minimize *EL*?
  - We free to assign x to either R_1 or R_2
  - Assigning x to region with smallest $L_{ij}p(x, C_i)$ will make *EL* smaller
- →Rule:
  - $L_{21}p(x, C_1) > L_{12}p(x, C_2)$ →predict y as C_1

$$\boxed{\frac{p(C_1|x)}{p(C_2|x)} > \frac{L_{12}}{L_{21}} \rightarrow predict\ y\ as\ C_1}$$

- 0/1 Loss: classify to the class which is more probable!

Loss and decision

- Continuous targets: squared loss
  - Given a model $p(x, y)$, minimize

$$EL = \int L\left(y, \hat{Y}(x)\right) p(x, y) dx dy$$

- Using **square loss**, the optimal is posterior mean

$$\hat{Y}(x) = \int y p(y|x) dy$$

ROC curves

- Binary classification
- The choice of the thershold $\hat{x} = \frac{L_{12}}{L_{21}}$ affects prediction→ what if we don't know the loss? Which classifier is better?
- **Confusion matrix**

	PREDICTED			
TRUE		1	0	Total
	1	TP	FN	N_+
	0	FP	TN	N_-

ROC curves

- **True Positive Rates** (TPR) = **sensitivity** = **recall**
  - Probability of detection of positives: TPR=1 positives are correctly detected

$$TPR = TP/N_+$$

- **False Positive Rates** (FPR)
  - Probability of false alarm: system alarms (1) when nothing happens (true=0)

$$FPR = FP/N_-$$

- **Specificity**

$$Specificity = 1 - FPR$$

- **Precision**

$$Precision = \frac{TP}{TP + FP}$$

ROC curves

- **ROC**=Receiver operating characteristics
- Use various thresholds, measure TPR and FPR
- Same FPR, higher TPR→ better classifier
- Best classifier = greatest Area Under Curve (**AUC**)