An Introduction to
Online Learning for Bayesians

Tim van Erven
An Introduction to Online Learning for Bayesians

Tim van Erven
Online Learning

- Decision problem
- Model: repeated game against an adversary
- Applications:
  - spam detection
  - data compression
  - online convex optimization
  - predicting electricity consumption
  - predicting air pollution levels
  - ...
Outline

- Online Learning
  - Introduction
  - Classification example
  - What can we achieve?
- Bayesian Methods
Repeated Game (Informally)

- Sequentially predict outcomes \( x_1, x_2, \ldots \)
- Measure quality of prediction \( a_t \) by loss \( \ell(x_t, a_t) \)
- Before predicting \( x_t \), get predictions (=advice) from \( K \) experts
- Goal: to predict as well as the best expert over \( T \) rounds.
- Data and Advice can be adversarial
Repeated Game

- Every round $t = 1, 2, \ldots$:
  1. Get expert predictions $a_t^k$ ($k = 1, \ldots, K$)
  2. Predict $a_t^*$
  3. Outcome $x_t$ is revealed
  4. Measure nonnegative losses $\ell(x_t, a_t^*), \ell(x_t, a_t^k)$

- Goal: minimize regret

$$\sum_{t=1}^{T} \ell(x_t, a_t^*) - \min_k \sum_{t=1}^{T} \ell(x_t, a_t^k)$$
Repeated Game

• Every round $t = 1, 2, \ldots$:
  1. Get expert predictions $a^k_t$ ($k = 1, \ldots, K$)
  2. Predict $a^*_t$
  3. Outcome $x_t$ is revealed
  4. Measure nonnegative losses $\ell(x_t, a^*_t), \ell(x_t, a^k_t)$

• Goal: minimize regret

\[
\sum_{t=1}^{T} \ell(x_t, a^*_t) - \min_k \sum_{t=1}^{T} \ell(x_t, a^k_t)
\]
Outline

- Online Learning
  - Introduction
  - Classification example
  - What can we achieve?
- Bayesian Methods
### Example: Spam Detection

<table>
<thead>
<tr>
<th>Subject</th>
<th>From</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gratis Turkije. . .</td>
<td>Reizen Center</td>
<td>$x_1 = 1$</td>
</tr>
<tr>
<td>uitnodiging hoorzitting reorganisatie FEW dinsdag 20 se...</td>
<td>Ivo van Stokkum</td>
<td>$x_2 = 0$</td>
</tr>
<tr>
<td>Re: Urgent Business Inquiry.</td>
<td>Ubc Ltd</td>
<td>$x_3 = 1$</td>
</tr>
<tr>
<td>Reminder: first colloquium</td>
<td>Jeu, R.M.H. de</td>
<td>$x_4 = 0$</td>
</tr>
<tr>
<td>Informatie over VUnet</td>
<td>College van Bestuur</td>
<td>$x_5 = 0$</td>
</tr>
<tr>
<td>USD 500 Free Deposit at PartyPoker!</td>
<td>PartyPoker</td>
<td>$x_6 = 1$</td>
</tr>
<tr>
<td>YOU ARE A WINNER!!! VERY URGENT NOTIFICATION.</td>
<td>UK INTL. LOTTERY PROMOTION</td>
<td>$x_7 = 1$</td>
</tr>
<tr>
<td>bachelor/master diploma uitreiking 14 september</td>
<td>Sotiriou, M.</td>
<td>$x_8 = 0$</td>
</tr>
<tr>
<td>HAPPY NEW YEAR 2068</td>
<td>Anil Shilpakar</td>
<td>$x_9 = 1$</td>
</tr>
<tr>
<td>Thailand Package</td>
<td>Anil Shilpakar</td>
<td>$x_{10} = 1$</td>
</tr>
</tbody>
</table>
Example: Spam Detection

- Labels: $x_t \in \{0, 1\}
- Predictions: $a_t \in \{0, 1\}
- 0/1-Loss: 
  \[ \ell(x_t, a_t) = \begin{cases} 
  0 & \text{if } a_t = x_t \\
  1 & \text{if } a_t \neq x_t 
  \end{cases} \]
- Experts: $K$ spam detection algorithms
- Regret: extra mistakes over best algorithm

\[
\sum_{t=1}^{T} \ell(x_t, a_t^*) - \min_{k} \sum_{t=1}^{T} \ell(x_t, a_t^k)
\]
Outline

• Online Learning
  • Introduction
  • Classification example
  • What can we achieve?

• Bayesian Methods
A First Algorithm

- Suppose one of the spam detectors is perfect

- Keep track of experts without mistakes so far:
  \[ S_t = \{ k \mid \text{expert } k \text{ made no mistakes before round } t \} \]

- Halving algorithm:
  \[ a_t^* = \text{majority vote among experts in } S_t \]

- Theorem: \( \text{regret} \leq \log_2 K \)
A First Algorithm: Halving

**Theorem:** regret \( \leq \log_2 K \)

- Does not grow with \( T \)

**Proof:**
- Suppose halving makes \( m \) mistakes, regret = \( m - 0 \)
- Every mistake eliminates at least half of \( S_t \)
- \( m \) is at most \( \log_2 |S_1| = \log_2 K \) mistakes
No Assumptions?

- Consider two trivial spam detectors (experts):
  \[ a_t^1 = 0 \quad a_t^2 = 1 \]
- I could be wrong all the time: \( x_t \neq a_t^* \)

Regret:
- Let \( n \) denote the number of ones in \( x_1, \ldots, x_T \)
- Total loss best expert: \( L := \min\{n, T - n\} \leq T/2 \)
- **Linear regret** = \( T - L \geq T/2 \)
Solution

- **Labels:** $x_t \in \{0, 1\}$
- **Predict probability** $a_t \in [0, 1]$ **that** $x_t = 1$
- **Expected 0/1-loss = absolute loss:**
  \[
  \ell(x_t, a_t) = |x_t - a_t|
  \]
- **Achievable regret:** $\sqrt{\frac{T}{2} \log K}$
- **$O(\sqrt{T})$ is standard in online learning**
Outline

- Online Learning
  - Introduction
  - Classification example
  - What can we achieve?
- Bayesian Methods
  - Data compression example
  - Mixable losses (or how to lie to Bayes)
  - Classification
  - When the posterior converges quickly...
Example: Data Compression

Big file

Small file

- Experts: $K$ data compression algorithms
- Regret: extra number of bits over best algorithm
Reduction to Online Learning

Data compression:

- \( x_1, \ldots, x_T \) are characters in original big file
- Can encode \( x_t \) using \(- \log P_t(x_t)\) bits, where \( P_t \) is a probability distribution I need to chose before seeing \( x_t \)

Online learning:
- Predict distribution \( P_t \) for \( x_t \)
- log loss: \( \ell(x_t, P_t) = - \log P_t(x_t) \)
Can We Guess the Regret?

- $K$ data compression algorithms
- For data compression I could use a two-part code
  1. $\log K$ bits identifies the best algorithm
  2. Concatenate with output of best algorithm
- **Regret:** $\log K$

- But in online learning I cannot split my output into two parts...
Bayes

- Experts define likelihoods:

\[ P(x_t \mid x_{1:(t-1)}, k) := P_t^k(x_t) \]

- Prior $\pi$ on unknown parameter $k \in \{1, \ldots, K\}$
Bayes

- Experts define likelihoods:

\[ P(x_t \mid x_{1:(t-1)}, k) := P^k_t(x_t) \]

- Prior \( \pi \) on unknown parameter \( k \in \{1, \ldots, K\} \)

\[ P^*(x_t \mid x_{1:(t-1)}) = \sum_k P(x_t \mid x_{1:(t-1)}, k) \pi(k \mid x_{1:(t-1)}) \]

where \( \pi(k \mid x_{1:(t-1)}) \propto P(x_{1:(t-1)} \mid k) \pi(k) \) is the posterior distribution on experts
Bayesian Regret

- Mix expert predictions according to their posterior probability

**Theorem:** If $\hat{k}$ is the best expert, then the Bayesian regret for log loss is at most $-\log \pi(\hat{k})$

- For uniform prior $\pi(k) = 1/K$ this is $\log K$, as expected.
- This is optimal as $K, T \to \infty$
Bayesian Regret

**Theorem:** If \( \hat{k} \) is the best expert, then the Bayesian regret for log loss is at most \(- \log \pi(\hat{k})\)

**Proof:**

- **Total loss:** \( \sum_{t=1}^{T} - \log P^*(x_t|x_{1:(t-1)}) = - \log P^*(x_{1:T}) \)
- **Marginal likelihood** \( P^*(x_{1:T}) \) is bounded by
  \[
P^*(x_{1:T}) = \sum_{k} P(x_{1:T} | k) \pi(k) \geq P(x_{1:T} | \hat{k}) \pi(\hat{k})
\]
- **Take negative logarithms**
- **Loss of best expert equals** \(- \log P(x_{1:T} | \hat{k})\)
Outline

- Online Learning
  - Introduction
  - Classification example
  - What can we achieve?
- Bayesian Methods
  - Data compression example
  - Mixable losses (or how to lie to Bayes)
  - Classification
  - When the posterior converges quickly...
How to Lie to Bayes

Log loss:

- **Likelihoods** \( P(x_t|x_{1:(t-1)}, k) = P_t^k(x_t) = e^{-\ell_{\log}(x_t, P_t^k)} \)
- **Loss is** \( \ell_{\log}(x_t, P_t) = -\log P_t(x_t) \)
How to Lie to Bayes

Log loss:

- Likelihoods $P(x_t | x_{1:(t-1)}, k) = P_t^k(x_t) = e^{-\ell_{\log}(x_t, P_t^k)}$
- Loss is $\ell_{\log}(x_t, P_t) = -\log P_t(x_t)$

General loss ("exponential weights"):

- Fix $\eta > 0$. Fake likelihoods

$$P(x_t \mid x_{1:(t-1)}, k) = e^{-\eta \ell(x_t, a_t^k)}$$

- Log loss equals $-\log P(x_t | x_{1:(t-1)}, k) = \eta \ell(x_t, a_t^k)$
How to Lie to Bayes

Log loss:

- Likelihoods $P(x_t | x_{1:(t-1)}, k) = P_t^k(x_t) = e^{-\ell_{log}(x_t, P_t^k)}$
- Loss is $\ell_{log}(x_t, P_t) = -\log P_t(x_t)$

General loss (“exponential weights”):

- Fix $\eta > 0$. Fake likelihoods
  \[
P(x_t | x_{1:(t-1)}, k) = e^{-\eta \ell(x_t, a_t^k)}\]
- Log loss equals $-\log P(x_t | x_{1:(t-1)}, k) = \eta \ell(x_t, a_t^k)$

These are not probabilities!
How to Lie to Bayes

Log loss:

- Likelihoods
  \[ P(x_t | x_{1:(t-1)}, k) = P_t^k(x_t) = e^{-\ell \log(x_t, P_t^k)} \]
- Log
  \[ \log(x_t, P_t) = -\log P_t(x_t) \]

But their values are in [0,1], so you cannot see that!

- Fix \( \eta > 0 \). Fake likelihoods
  \[ P(x_t | x_{1:(t-1)}, k) = e^{-\eta \ell(x_t, a_t^k)} \]

- Log loss equals
  \[ -\log P(x_t | x_{1:(t-1)}, k) = \eta \ell(x_t, a_t^k) \]
Mixability

If the loss is not log loss and predictions are not probabilities, then you cannot predict with the posterior distribution

$$P^*(x_t|x_{1:(t-1)}) = \sum_k P(x_t|x_{1:(t-1)}, k) \pi(k|x_{1:(t-1)})$$
If the loss is not log loss and predictions are not probabilities, then you cannot predict with the posterior distribution

\[ P^*_t(x_t|x_{1:(t-1)}) = \sum_k P(x_t|x_{1:(t-1)}, k) \pi(k|x_{1:(t-1)}) \]

I only need **mixability**...

A loss is \( \eta \)-**mixable** if, for any posterior distribution, we can find a prediction \( a^* \) that is at least as good:

\[
e^{-\eta \ell(x_t, a^*)} \geq P^*_t(x_t|x_{1:(t-1)}) \quad \text{for any } x_t
\]
Mixability

If the loss is not log loss and predictions are not probabilities, then you cannot predict with the posterior distribution

\[ P^* (x_t | x_1:(t-1)) = \sum_k P(x_t | x_1:(t-1), k) \pi(k | x_1:(t-1)) \]

I only need **mixability**...

A loss is $\eta$-mixable if, for any posterior distribution, we can find a prediction $a^*$ that is at least as good:

\[ e^{-\eta \ell(x_t, a^*)} \geq \sum_k P(x_t | x_1:(t-1), k) \pi(k | x_1:(t-1)) \quad \text{for any } x_t \]
Mixability

If the loss is not log loss and predictions are not probabilities, then you cannot predict with the posterior distribution:

$$P^*(x_t|x_{1:(t-1)}) = \sum_k P(x_t|x_{1:(t-1)}, k)\pi(k|x_{1:(t-1)})$$

I only need **mixability**...

A loss is \(\eta\)-mixable if, for any posterior distribution, we can find a prediction \(\alpha^*\) that is at least as good:

$$e^{-\eta \ell(x_t, \alpha^*)} \geq \sum_k e^{-\eta \ell(x_t, \alpha_t^k)} \pi(k|x_{1:(t-1)})$$

for any \(x_t\).
Mixability

If the loss is not log loss and predictions are not probabilities, then you cannot predict with the posterior distribution:

\[ P^*(x_t|x_{1:(t-1)}) = \sum_k P(x_t|x_{1:(t-1)}, k) \pi(k|x_{1:(t-1)}) \]

I only need **mixability**...

A loss is \(\eta\)-mixable if, for any distribution \(w(a)\), we can find a prediction \(a^*\) that is at least as good:

\[ e^{-\eta \ell(x,a^*)} \geq \sum_a e^{-\eta \ell(x,a)} w(a) \quad \text{for any } x \]
Mixable Losses

• Regret bounded by \(-\frac{\log \pi(\hat{k})}{\eta}\)

• For largest possible \(\eta\) this is optimal as \(K, T \rightarrow \infty\)

Examples:

• Square loss is 2-mixable:

\[
\ell(x_t, a_t) = (x_t - a_t)^2 \quad x_t, a_t \in [0, 1]
\]

• Relative entropy loss is 1-mixable:

\[
\ell(x_t, a_t) = x_t \log \frac{x_t}{a_t} + (1 - x_t) \log \frac{1 - x_t}{1 - a_t} \quad x_t, a_t \in [0, 1]
\]

• Absolute loss is not \(\eta\)-mixable for any \(\eta > 0\)
Mixable losses

Theorem 1: The Bayesian regret for log loss is at most $- \log \pi(\hat{k})$

Theorem 2: The Bayesian regret for any $\eta$-mixable loss is at most $\frac{- \log \pi(\hat{k})}{\eta}$

Proof by reduction to log loss:

$$\sum_{t=1}^{T} \eta \ell(x_t, a_t^*) - \min_k \sum_{t=1}^{T} \eta \ell(x_t, a_t^k) \leq \sum_{t=1}^{T} \ell_{log}(x_t, P(\cdot|a_t^*)) - \min_k \sum_{t=1}^{T} \ell_{log}(x_t, P_t(\cdot|a_t^k)) \leq - \log \pi(\hat{k})$$
Log Loss is Special

- Reduction to log loss suggests that:
  “All mixable losses are like log loss in some way”

- New characterization of mixable losses captures in which way. [vE, Reid, Williamson, 2011]
Outline

- Online Learning
  - Introduction
  - Classification example
  - What can we achieve?
- Bayesian Methods
  - Data compression example
  - Mixable losses (or how to lie to Bayes)
  - Classification
  - When the posterior converges quickly...
Absolute Loss

- **Labels**: $x_t \in \{0, 1\}$
- **Predict probability** $a_t \in [0, 1]$ such that $x_t = 1$
- **Expected 0/1-loss = absolute loss**:

$$\ell(x_t, a_t) = |x_t - a_t|$$
Absolute Loss

- **Labels:** \( x_t \in \{0, 1\} \)
- **Predict probability** \( a_t \in [0, 1] \) **that** \( x_t = 1 \)
- **Expected 0/1-loss = absolute loss:**

\[
\ell(x_t, a_t) = |x_t - a_t|
\]

- **Not mixable...**
- **But can be approximated by an** \( \eta \)-**mixable loss up to approximation error** \( \frac{\eta}{8} \) **per round!**
Bayes for Absolute Loss

**Theorem:** Bayes for absolute loss with
\[
\eta = \sqrt{\frac{8 \log K}{T}}
\]
has regret at most
\[
\sqrt{\frac{T}{2} \log K}
\]

**Proof:**

- If loss were mixable, the regret would be bounded by \( \frac{\log K}{\eta} \)
- Approximation error: \( \eta/8 \) per round
- Resulting bound:
  \[
  \frac{\log K}{\eta} + \frac{\eta T}{8}
  \]
Outline

• Online Learning
  • Introduction
  • Classification example
  • What can we achieve?
• Bayesian Methods
  • Data compression example
  • Mixable losses (or how to lie to Bayes)
  • Classification
  • When the posterior converges quickly...
Converging Posterior

- Approximation error $\frac{n}{8}$ does not depend on the posterior distribution
- If the posterior distribution converges we can do better...
Converging Posterior

- Approximation error $\frac{\eta}{\delta}$ does not depend on the posterior distribution.
- If the posterior distribution converges we can do better...

Lemma: For $\eta \leq 1$ the approximation error is bounded by

$$(e - 2)\eta \left(1 - \pi(k \mid x_{1:(t-1)})\right)$$

for any $k$. [vE, Grünwald, Koolen, De Rooij, 2011]
Converging Posterior

- Can choose $\eta$ such that the regret is bounded by:

1. If the posterior converges sufficiently fast:
   
   $O(K)$

2. Always, even if the posterior does not converge:
   
   $O(\sqrt{T \log K})$
Outline

- Online Learning
  - Introduction
  - Classification example
  - What can we achieve?
- Bayesian Methods
  - Data compression example
  - Mixable losses (or how to lie to Bayes)
  - Classification
  - When the posterior converges quickly...
Summary

- **Online Learning**
  - Repeated prediction game
  - Examples: data compression, classification
  - Want sublinear regret: constant or $O(\sqrt{T})$

- **Bayesian Methods**
  - Generalization to mixable losses
  - Generalization to classification
  - Better classification when posterior converges quickly
Online Learning

Prediction with Expert Advice:
• Finite/countable number of experts

Online Convex Optimization:
• Learn convex combinations of experts
Online Learning

Prediction with Expert Advice:
- Finite/countable number of experts

Online Convex Optimization:
- Learn convex combinations of experts

Gradient trick:
replace a convex loss by a linear approximation
References

- **Standard textbook:**

- **Course slides by Peter Bartlett:**

- **Van Erven, Reid and Williamson.** *Mixability is Bayes Risk Curvature Relative to Log Loss.* COLT 2011.

- **Van Erven, Grünwald, Koolen and De Rooij.** *Adaptive Hedge.* NIPS 2011.

- **De Rooij, Van Erven, Grünwald, Koolen.** *Follow the Leader If You Can, Hedge If You Must.* Submitted, 2013.